THE UNIVERSITY of ADELAIDE

School of Economics

Working Papers

ISSN 2203-6024

Counter Intuitive Learning: An Exploratory Study

Nobuyuki Hanaki
Alan Kirman
Paul Pezanis-Christou

Working Paper No. 2016-12
July 2016

Counter intuitive learning: An exploratory study[*]

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

July 31, 2016

Abstract

The literature on learning in unknown environments emphasises reinforcing on actions which produce positive results. But, in some cases, success requires shifting from a currently successful actions to others. We examine, experimentally and theoretically in a very simple framework, how individuals initially learn by exploiting information from the pay-offs of actions taken but also from exploring new actions. We analyse if and how they learn that pay-offs are inter-temporally dependent. We then ran the same experiments but where individuals could observe the actions taken or the pay-offs obtained by others or both. Such observations improved pay-offs if one of the pair had learned to obtain the maximum pay-off.

Keywords: multi-armed bandit, reinforcement learning, eureka moment, pay-off patterns, observational learning

JEL Code: D81,D83

[*] We thank Ido Erev, Mamoru Kaneko, Rajiv Sethi and audiences at SAET 2014 (Tokyo), ASSET 2014 (Aix-en-Provence), and the Universities of Nice (GREDEG) and Toulouse for useful comments and suggestions. Support from the School of Economics of the University of New South Wales where the experiments were conducted, LABEX OT-MED (ANR-11-LABX-0061), a JSPS-ANR bilateral research grant "BECOA" (ANR-11-FRJA-0002) and the Australian Research Council through a *Discovery Project* grant (DP140102949) are gratefully acknowledged. The usual disclaimer applies.

[†] Université Côte d'Azur, SKEMA, CNRS, GREDEG, and IUF.

[‡] CAMS, EHESS, and Aix Marseille University. [Corresponding author: alan.kirman@univ-amu.fr]

[§] School of Economics, University of Adelaide.

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 that there are other options available to them as they proceed.

Most of the emphasis in the learning literature is on reinforcing 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). However, they 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 in the literature as the "exploration, exploitation" trade-off (March, 1991) and we will use this as our basic framework. We will assume that agents start with a model of their environment and try to learn within it.

But, what if this model is an inadequate representation? Can agents learn not only which actions to take in their currently perceived model but also to change the latter? In our theoretical model, we will model the process involved in doing this.

To focus the discussion, consider the case in which the success of an endeavor depends on shifting from a currently successful action to another. 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. The use of such inter-temporal strategies depends on the farmer becoming aware of the existence of 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, behavioral strategies is not yet very well understood. How individuals learn which actions to take within their perceived model has been thoroughly investigated but how they learn to change their perception of their model is much less clear.

This raises an important question which has generated a large literature. What are the mental processes at work that lead individuals to learn how to understand and operate in their environment? This has been the subject of considerable interest in the neurosciences where attention has been focused on the mechanisms which lead a person to start exploring new alternatives and thus to enlarge his vision of the world he lives in.

One approach has been to identify the neural processes at work when an individual suddenly perceives the answer to a problem he or she has been wrestling with. This is often referred to as the “aha” or, referring to Archimedes, the “eureka” moment. 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.

We are concerned with a more general situation in which individuals take actions and try to improve their performance but do not know whether there is necessarily a “solution.” What interests us, in particular, is how agents come to decide to try alternative actions to those which they have already used. It is now argued, in the neuroscience literature, 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 Laureiro-Martínez et al., 2014, 2015).[1]

[1] 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

In this paper, we focus on experiments to investigate how individuals gain a better understanding of their environment and, as a result, obtain better results. We study a case in which the payoffs from actions are correlated and have a temporal structure. More specifically, we study a case in which subjects are faced with a multi-armed bandit and try to obtain the highest pay-off they can get from choosing the different arms in succession. Our experiment is very different from other experiments on "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 that framework, agents will optimally try arms to form an idea as to their expected pay-offs 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. Such a strategy is not well adapted to a framework like ours in which payoffs have temporal structure and are not independent of each other.

How individuals learn which action to take, in the sort of context that corresponds to our experimental framework, and how to condition that choice on previous experience has been a subject of considerable interest in other fields such as machine learning. There, a problem which corresponds to our simplest case, is what is referred to as the "contextual bandit problem." 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. The strongest known results (Auer, 2002; McMahan and Streeter, 2009; Beygelzimer et al., 2011) provide algorithms that carefully control the distribution of the probabilities with which actions are explored, to achieve an optimal regret after T rounds. However, in this paper, we will use a rather simple exploratory rule, "entropic" exploration. The logistic rule which we employ can be derived as an optimal trade-off between exploitation and exploration (Nadal et al., 1998).

this process is also present in longer term situations.

An important feature of our problem, as mentioned, is the sequential nature of the pay-offs and, until the subject recognises this structure, he may continue to try to reinforce on the experience from choosing single arms. The important change that is necessary for him to learn how to improve his gain, is to recognise that there is a sequential pattern in the pay-offs and then reinforce on those. Thus he has to change the space of possible choices to one of choices conditioned on what has happened in the past, and this, in our framework, is the "context" which is referred to in the machine learning literature to which we have just alluded. Indeed, the literature on machine learning has dealt with similar issues when analyzing problems in "sequence extrapolation" (Laird and Saul, 1994). That literature typically tries to develop deterministic algorithms that may be used for successful sequence predictions in general. And where there is an underlying deterministic process generating pay-offs as in our case, such an approach is appropriate.

Our paper provides a link between two approaches, that of machine learning in which the recognition of patterns in the data being observed is the primary objective, and that developed in game theory in which the objective is to recognise patterns in opponents' behaviour (see, Sonsino, 1997; Spiliopoulos, 2012, 2013). Although the mechanics of the processes have many aspects in common, the goal is very different. In machine learning, the aim is to recognise and identify patterns in data. In the game theory literature, the objective is to react to the patterns in the best way knowing that the reaction will provoke a change in the patterns. The ultimate aim is to find an equilibrium in which the patterns no longer change.

We will proceed by first presenting our experimental framework in which the subjects have to understand that the order in which they choose actions is important if they are to achieve good results. Until subjects realise that there is such a sequential structure, they cannot obtain a high or, although they do not know it, the optimal pay-off. Their problem is precisely that they have no a priori information about how actions and pay-offs are linked. Our question is not whether they can find the solution to a well defined problem, such as solving the Rubik's cube, but rather, whether, by exploring the various possibilities and using the information they gain in doing so, they improve their payoffs and may even arrive, without realising it, at the maximum attainable. We present some examples of how subjects performed in the experiments and then develop a model to try to capture how subjects learn about the interdependence in the pay-offs from their actions and how they learn to try new strategies. We then compare the behaviours the model generates with those observed in the experiments.

In the second part of the paper, we consider another important aspect of the learning problem. In many cases, individuals do not learn in isolation but observe what others do or what they gain from their actions. In a second series of experiments, we took this into account and allowed subjects to observe another individual's choices, pay-offs or both. There is a substantial literature on what is referred to as "observational learning" (Bandura and McDonald, 1963; Bandura et al., 1963; Fryling et al., 2011) and we investigate to what extent individuals benefit from such information by accelerating their learning process.

2 The baseline experiment

In our baseline experiment, 30 subjects participated in an experiment that involve 200 rounds of play. In each of these round, participants were asked to choose one of four options without being given any information about the possible payoffs that each option 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 $Mod(t-1, 3) = 3 - a$ and 0 otherwise. 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 option choice cycles in the order of 3, 2, 1, 3, 2, 1, 3,... from round 1, ..., 200.

At the end of each round, participants received no other information than the payoff outcome of their own choice. Further, they were not allowed to record any information on paper or else either, so that they had to rely on their memory of past play and outcomes to make their decisions. To incentivize their decisions, their payoffs (expressed in Experimental Currency Units, ECUs) were cumulated over the 200 rounds and converted into cash at the end of the experiment at the rate of 20 Australian cents per ECU.

Once the 200 rounds of play were over, participants were given a new set of instructions that outlined the lottery choice questionnaire to measure their risk preferences proposed by Holt and Laury (2002). Again, to incentivize decisions, it was explained to them that, once the questionnaire was completed, one of the lottery questions would be randomly chosen to determine their individual payoffs for this second part of the experiment. Subjects received the outcome of the lottery if they accepted the chosen lottery and nothing if they did not. The lottery payoffs, which were expressed in Australian Dollars, are provided in the Appendix along with the full set instructions.

The average individual reward from participating in this experiment, including the lottery choice questionnaire, was A$30 (which includes 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.

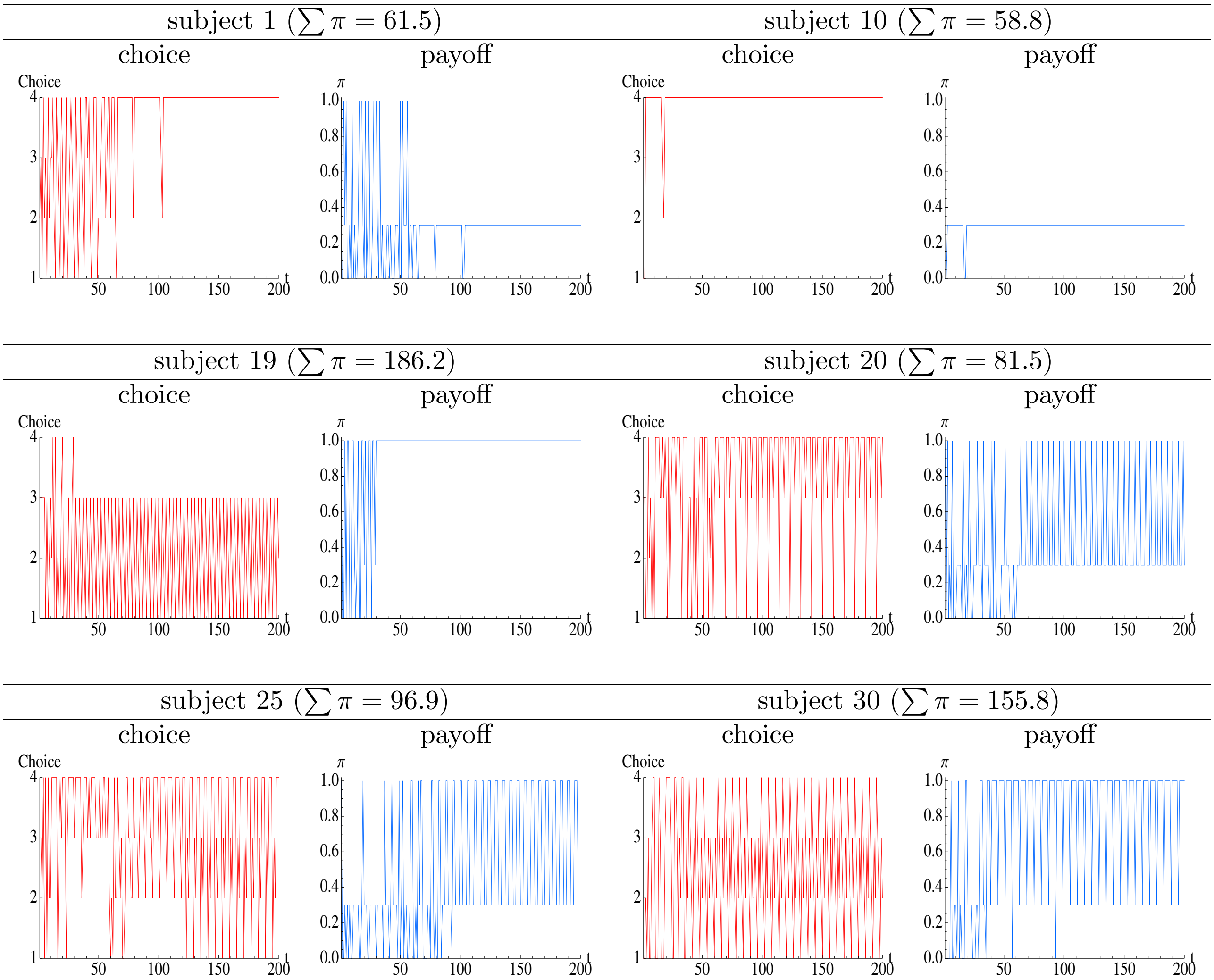
We are interested in understanding how quickly subjects discover, if ever, that there was a sequential structure to the pay-offs, and how many actually made the best choices given the true hidden payoff generating pattern even though they were not aware of the existence of such a pattern. In other words, how many subjects "learned to optimise."

2.1 Descriptive results

We start with providing a brief model-free overview of the most salient types of behaviour. Out of the 30 subjects, exactly half of them discovered the hidden cycle that generates a payoff of 1 in each round but they did so at very different times: five of them found it within 50 rounds, 13 within 100 rounds and 15 within 200 rounds. The other half of subjects either never found it or settled for the safe option (option 4) which generates a constant payoff of 0.3.[2]

Figure 1 shows the time-series of choices and payoff outcomes for six participants who displayed different types of behaviour. While Subject 19 discovered the hidden cycle within 50 rounds, other subjects explored for a few rounds and settled for the safe option. Among the latter, some never got a payoff of 1 when choosing other options (e.g., Subject 10) while others decided to exploit the safe choice even though they sometimes got a payoff of 1 when choosing other options (e.g.,

[2] By "discover", we mean that a subject uses the "best" pattern and earn a payoff of 1 for successive rounds. We define a participant as having discovered the best pattern if he or she earned a payoff of 1 for twelve consecutive rounds for the first time and used the middle round of this sequence as the discovery time.



subject 1 ($\sum \pi = 61.5$)

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

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

subject 20 ($\sum \pi = 81.5$)

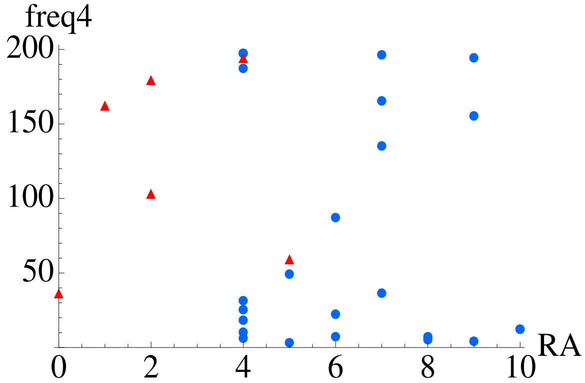
subject 25 ($\sum \pi = 96.9$)

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

choice

payoff

Figure 1: Exemples 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.

Subject 1). Yet, other participants settled into more complex and less profitable cyclical choice patterns: Subject 20 settled into a 9-round cycle repeating the choice sequence (3, 4, 4, 4, 4, 1, 4, 4, 4), Subject 25 settled into the 6-round cycle (1, 3, 2, 4, 4, 4) while Subject 30 settled into another, almost optimal, 6-round cycle (3, 2, 1, 3, 2, 4).

One might reasonably conjecture that more 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 to 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 riskier than the other in the early tasks and safer in the later ones. A risk averse subject should therefore choose the safer lottery in the early tasks and switch to the other lottery in the later ones; the later this switch occurs and the more risk averse is the subject.[3]

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

[3] 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 uncertainty, 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.

the reported coefficient is not statistically significant at $\alpha = 0.05$, we discard risk aversion as a possible explanation for the observed behaviour and develop a model that attempts to capture the observed behaviour.

3 The baseline 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 if it is observed once. 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 period t. Thus, if agent i has experienced all the possible outcomes from choosing all four options at least once, before period 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. In addition, we need to decide how the agents in our model make their choices amongst these different options. To what extent should they rely on their experience in trying the different options and to what extent should they switch to trying on other options. This is the well known "Explore v.s. Exploit" trade-off which has given rise to an extensive literature (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. Or put alternatively, the agents are assumed to make that assumption. This, we assume, is their initial model of the environment. Now the question is, how do individuals learn come to understand that their model is an inadequate representation of reality and to change their model?[4]

[4] One approach is that of econometricians who argue that what is being learned are the characteristics of the data generating process with which agents are faced (see, Dufour, 2008, for a survey). However, whilst various forms of Bayesian learning have been explored, the problem of how to act when the actual data generating process does not fall into the class which the agent has considered, remains a difficult one. One approach is to start with a very general specification in which all possible variables, functional forms, lags etc are included in the set of candidate models. However, it has been argued that to start with such a general specification and then to narrow it down to a specific model is simply too costly. But Castle et al. (2013) give a full discussion of this problem and argue that it is now technically possible and not necessarily very costly if one uses automatic selection procedures. Nevertheless it seems

It may well be the case that individuals do not learn the "correct model." Indeed, 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 (see, e.g., Arrow and Green, 1973; Kirman, 1975, 1983; Bray, 1982; Woodford, 1990). In those cases the pay-offs are, themselves, modified by the actions taken by the economic agents. Here, this is not the case. But the agents may be led, depending on the pay-offs that they receive in the early rounds, to conclude that they should use the safe arm, for example. But we are faced with a much more difficult question as to what it is that individuals might observe that might lead them to change their model of the world. This would lead them to be faced with a different set of possible actions and possibly arrive at results quite different from those obtained in the past. One reason for making such a change would be that the results obtained using the original perception were consistently different from those anticipated. In many cases, such as those discussed in the literature just mentioned, this would not be evident. An individual may not understand that there were alternative courses of action which would be superior to the best of those that he has already tried. He "does not know what he does not know." However, our experimental evidence shows that many subjects do become aware that there is an inter-temporal structure in the way pay-offs are generated.

The first step will be for individuals to recognise that their gain does not simply depend on the particular action taken but also on when that action is taken 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 the "context." In our case, the context is the order in which actions are taken. A substantial literature on learning about "contextual bandits" has developed in machine learning. In the simplest case, corresponding to our baseline treatment, the feedback is incomplete: in any given round, the agent observes the reward only for the chosen action but is not informed as to what he would have obtained had he chosen another action. The use of features to encode context is inherited from supervised machine learning and in our case, when we come to consider other experimental treatments, this will allow us to introduce the history or sequence of previous choices

unlikely that this is the procedure that ordinary subjects would adopt.

and not just the pay-offs from each. On the other hand, that literature emphasises that exploration is necessary for good performance as it is in reinforcement learning.[5] In our particular case, the recognition that sequences matter amounts to detecting patterns in the payoff sequence and again in the machine learning literature considerable attention has been paid to pattern detection.[6]

For our model then, to take these considerations into account, 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. For example, if $l^i = 2$ and $r^i = 2$, after observing the sequence of choice payoff pair $(1, 1), (1, 0)$ over two consecutive periods twice, agent i will start considering the existence of a dependency (namely, choosing the option 1 after choosing it and obtaining payoff of 1 will result in payoff of 0).[7]

Once i starts considering such dependencies, he will start choosing the option conditional on the outcome in the previous period. In particular, i will start remembering the possible payoffs each conditional choice generates. Let $\Pi^i_t(a|h)$ represent the set of payoffs agent i has observed by choosing option a conditional on history h until period t. For example, $\Pi^i_t(1|(2, 1))$ 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 periods, we restrict our attention to those choices which are conditioned

[5] As Agarwal et al. (2014) point out, contextual bandit problems are found in many important applications such as online recommendation and clinical trials. The use of the term contextual is important because it suggests that other features of the choices can be taken into consideration, thus one can argue that this sort of model represents an intermediate point between "supervised learning" and pure "reinforcement learning."

[6] An even more difficult problem that has received some attention when dealing with games is that of recognising patterns in an opponent's behaviour. Here the difficulty is that the pay-off to each player depends on both players' choices. One attempt to do that was made by Sonsino (1997) and his idea of "learning to learn" in this case, is simply illustrated when he says, "As a 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." Recently, Spiliopoulos (2012, 2013) showed that introducing ability to recognize pattern in the opponent's choices in belief-based learning model greatly improves the fit of the model to the experimental data. In our case the subject can be thought of as playing "a game against nature" where nature is playing a deterministic sequence. We do not need to give such an interpretation however, and the problem boils down to one of identifying a structure in the pay-offs without considering that they are generated by an "opponent."

[7] 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.

only on the outcome of the most recent period (i.e., the choice of period t, a_t, depends on the outcome of period $t-1$ ($h^i_{t-1} = (a^i_{t-1}, \pi^i_{t-1})$)). 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 period, h^i_{t-1}, 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 period t choices of options conditional on the outcomes in period $t-1$, $s = a|h$. Agents start using conditional strategies only after they have become aware of existence of a sequential pattern as we have described above.

Let us first describe an unconditional strategy. $A^i_t(a)$ summarizes, at the beginning of period t, the past experience for agent i from choosing option a. Let $A^i_0(a) = 0.5$ for all i and a. We assume that $A^i_t(a)$ evolves as follow:

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

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

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

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

Let $B^i_t(a|h)$ summarize i's experience from choosing a conditional strategy $a|h$ at the beginning of period t. Let $B^i_t(a|h) = A^i_t(a)$ for all h after i became aware of the inter-temporal dependencies. And $B^i_t(a|h)$ evolves as follows

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

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

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

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

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

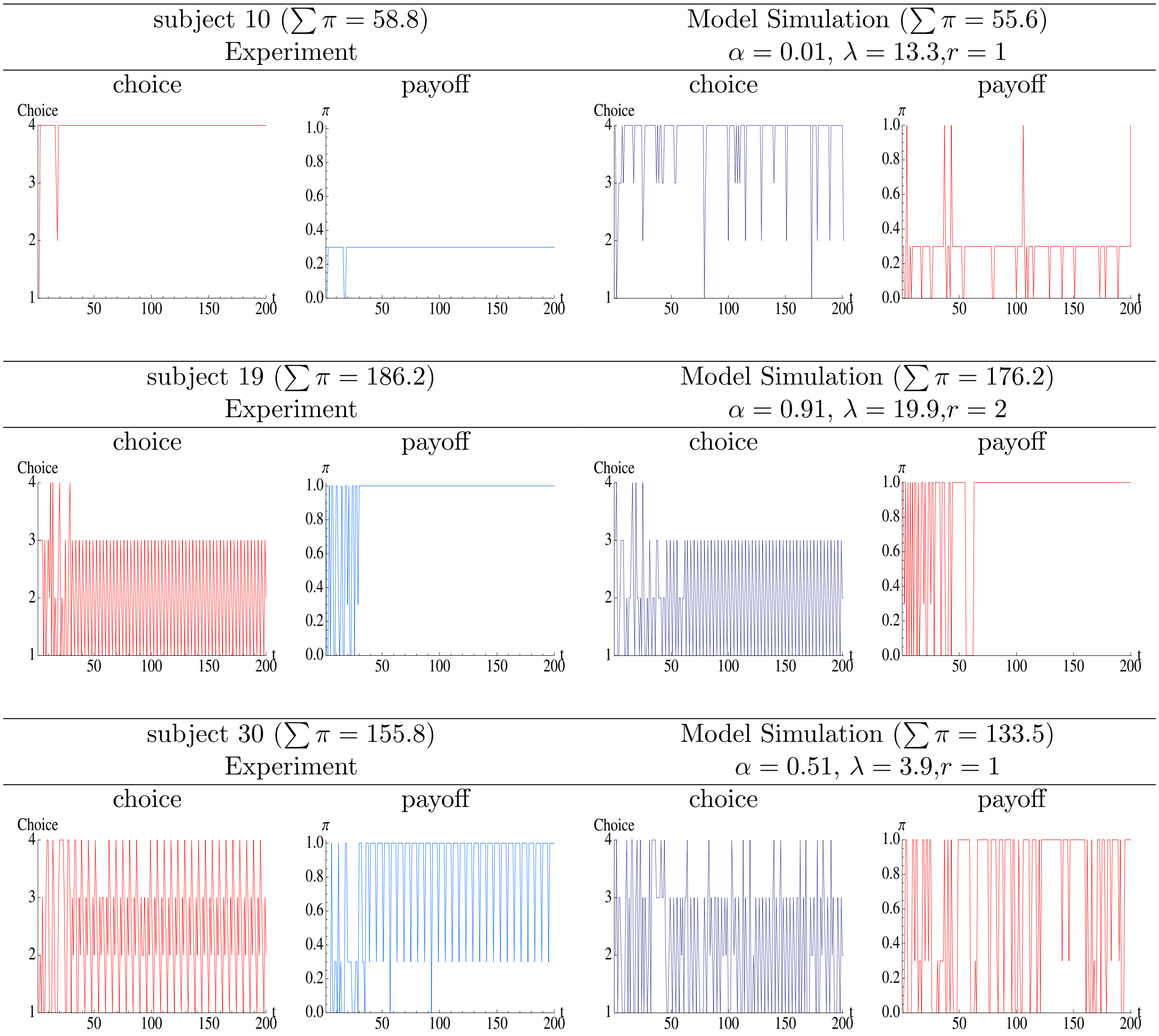
where $P_t^i(a_t^i)$ is the probability of observing choice a_t^i according to the model described above.

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 step of 0.1, and $r^i \in [1, 10]$ with a step of 1. When multiple sets of parameter values generate the same sum of log likelihood, we selected the one with the smallest r^i as this parameter was the unique source of multiplicity.

Figure 3 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).

Once we fitted and determined the set of best parameter values for each individual $\left\{(\hat{\alpha}^i, \hat{\lambda}^i, \hat{r}^i)\right\}$, we considered a population of 1000 artificial agents with parameters randomly drawn (with replacement) from this set and we compare the simulated outcomes to the observed ones in terms of individuals' discovery times and total payoffs. Clearly, as participants did not know the exact structure underlying the bandits' payoffs, their objective function is ill-defined and many choice patterns could be believed to be 'optimal'. Here we focus on the discovery of the hidden cycle generating a payoff of 1 in each round. More precisely, we diagnose a discovery if a participant earns a payoff of 1 for twelve consecutive rounds for the *first time*[8] and we define participant i's

[8] As already observed, some participants occasionally 'deviated' from the hidden pattern after having exploited it



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

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

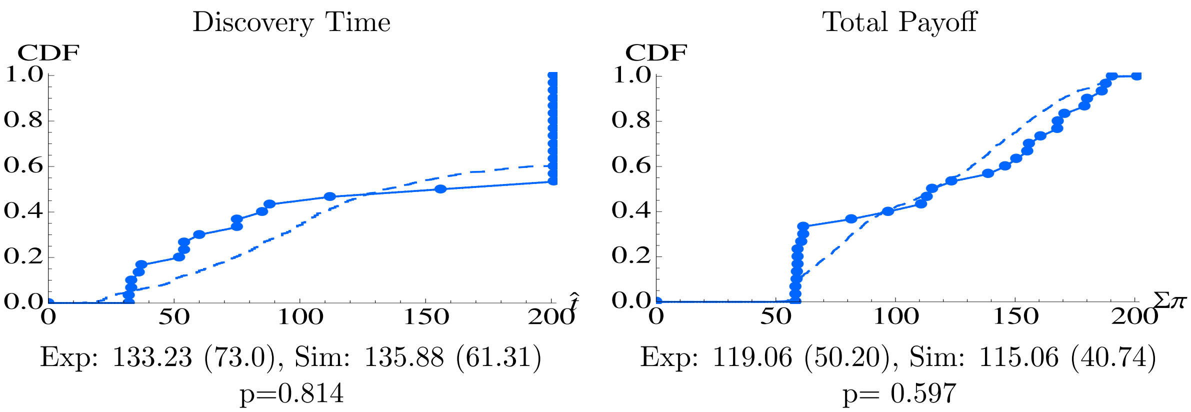
subject 19 ($\sum \pi = 186.2$) Experiment

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

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

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

Figure 3: Comparison of the experimental data and simulated model for three subjects.


Discovery Time

Total Payoff

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 4: CDFs of discovery times and total payoffs. Legend: Observed (solid), Simulated (dashed). Mean (standard deviation) are also shown. P-values are based on Permutation tests (for independent samples), two-tailed.

discovery time, $\hat{t}^i$ by the median round number of this twelve-round window.[9]

The observed and simulated cumulative distributions of discovery times and total payoffs and their respective averages (standard deviations) reported in Figure 4 indicate that the model fits these data remarkably well. This is confirmed by the Permutation tests (for independent samples) which do not reject the null of equivalent data samples ($p = 0.814$ for discovery times and $p = 0.597$ for total payoffs, two-tailed tests).

4 Additional experiments: Learning from others

We now extend our baseline setting, henceforth named the NI (No Information) treatment, to allow participants to observe the choices made or the payoffs obtained by another participant. This "observational learning," that is learning from others, came to the forefront with the work of Bandura and his colleagues (Bandura and McDonald, 1963; Bandura et al., 1963). They argued that the process of learning is greatly influenced by observing the behaviour of other individuals and Fryling et al. (2011) argue that "The important conclusion of these studies is that behavior

for long time, but most of them quickly returned to it.

[9] For example, if a subject got a payoff of 1 from round 40 until 200, then his her discovery time is round 45. Note that we get qualitatively the same results if we change the criterion to six consecutive rounds instead of twelve.

change can and does occur through observation, even when such observation is incidental, occurring in the context of other activities." Indeed, when individuals are looking at the other's behaviour or rewards what may cause them to modify their own behaviour is the observation of systematic patterns in the behaviour of the person they are observing. In our framework, the observation of consistently high pay-offs or of regular patterns in the choices may lead a subject to change his behaviour.[10] 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. They add, however, an important caveat: in a non-stationary environment, imitation may produce inertia in behavior as imitation induces herd behaviour and discourages individual exploration. In the experiments that we will discuss in this paper, the environment is, in fact, stationary, but the subjects were not aware that this was the case.

Imitation or social learning has also been taken into account in the neuroscientific field by conducting experiments in which subjects, in addition to their own experience also profit from feedback from other participants. It is also probably appropriate when considering phenomena such as the emergence of agricultural practices which involve substantial periods of time before the results of certain actions are known, and where actions are taken in a noisy environment, where weather and other stochastic natural variations can make the interpretation of the consequences of an action difficult. In such environments one might expect the learning process to be a collective one and still the adoption of innovative actions to be very slow. Indeed, in the optimal foraging literature in which individuals "learn" to behave optimally (see, Stephens and Krebs, 1986), no indication is given as to the time necessary for such convergence to optimal behaviour to be accomplished. Furthermore, one might argue that the process never converges since the environment may change more quickly than either individual or collective learning. In fact, one can either think of the problem of individuals learning how to solve a problem which is well defined even if they may not

[10] Similar work has been done by 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. Many subjects managed to make the pay-offs from the two arms equal but did not find the optimal behaviour. In their case observing the behaviour of others actually had negative effects in some cases. There is a considerable literature on observational learning that is relevant here (e.g. Smith and Sørensen, 2011). However, most (if not all) of it deals with interactive situations where pay-offs to individuals depend on those to others, which is different from what we do.

Table 1: Summary of experimental treatments

Treatment	No. of subjects	Description
NI	30	Each subject independently tries our bandit problem
CPI	32 (16 pairs)	Subject can observe both the choice made by one another subject and the payoff s/he obtained (the pair is fixed throughout the experiment).
PI	32 (16 pairs)	Subject can observe the payoff obtained by one another subject (the pair is fixed throughout the experiment).
CI	32 (16 pairs)	Subject can observe the choice made by one another subject (the pair is fixed throughout the experiment).

be aware of the full structure of the problem, or think of individuals trying constantly to adapt to an environment which is non-stationary.[11]

What we investigate here is the relative impact on subjects' behaviour of three different sorts of information: observing another player's choice and pay-off, observing another's choice only, and observing only the other's pay-off. To be more precise each of the following three experiments (treatments) involve 32 different subjects who were randomly matched in pairs for the 200 rounds of play. In one treatment, they got to know both the other subject's last choice and payoff outcome. In another treatment, they got to know the other subject's last payoff outcome but not the corresponding choice whereas in the last one, they got to know the other's last choice but not the payoff outcome of that choice. We will refer to these treatments as the CPI (Choice and Payoff Information), PI (Payoff Information) and CI (Choice Information) treatments, respectively. All other aspects of these treatments were the same as those of the NI treatment – Table 1 summarizes the four treatments' main features.

[11] It is worth noting that some work has been done in neuroscience to try to detect what mechanisms are at work when individuals shift to another way of looking at their environment, (see, Cohen et al., 2007) in which individuals were faced with a changing environment. Thus they were not faced with a problem to be solved but rather with a problem of adapting their choices to this evolving environment.

4.1 Descriptive Results

Figures 5, 6 and 7 display the time-series of choices and payoffs for four pairs of subjects in each of these three additional treatments, CPI, PI, and CI, respectively. 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 (Subjects 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, however, we find no evidence of participants settling into more complex and less profitable choice patterns.

Before proceeding with the modeling of behaviour, we check again whether the observed frequencies of safe choices were driven by the participants' risk aversion. The scatter plots in Figure 8 suggest no clear pattern in this regard, and since the reported correlation coefficients are not significant ($p > 0.06$) we reject the conjecture that behaviour in these information treatments is driven by risk aversion.

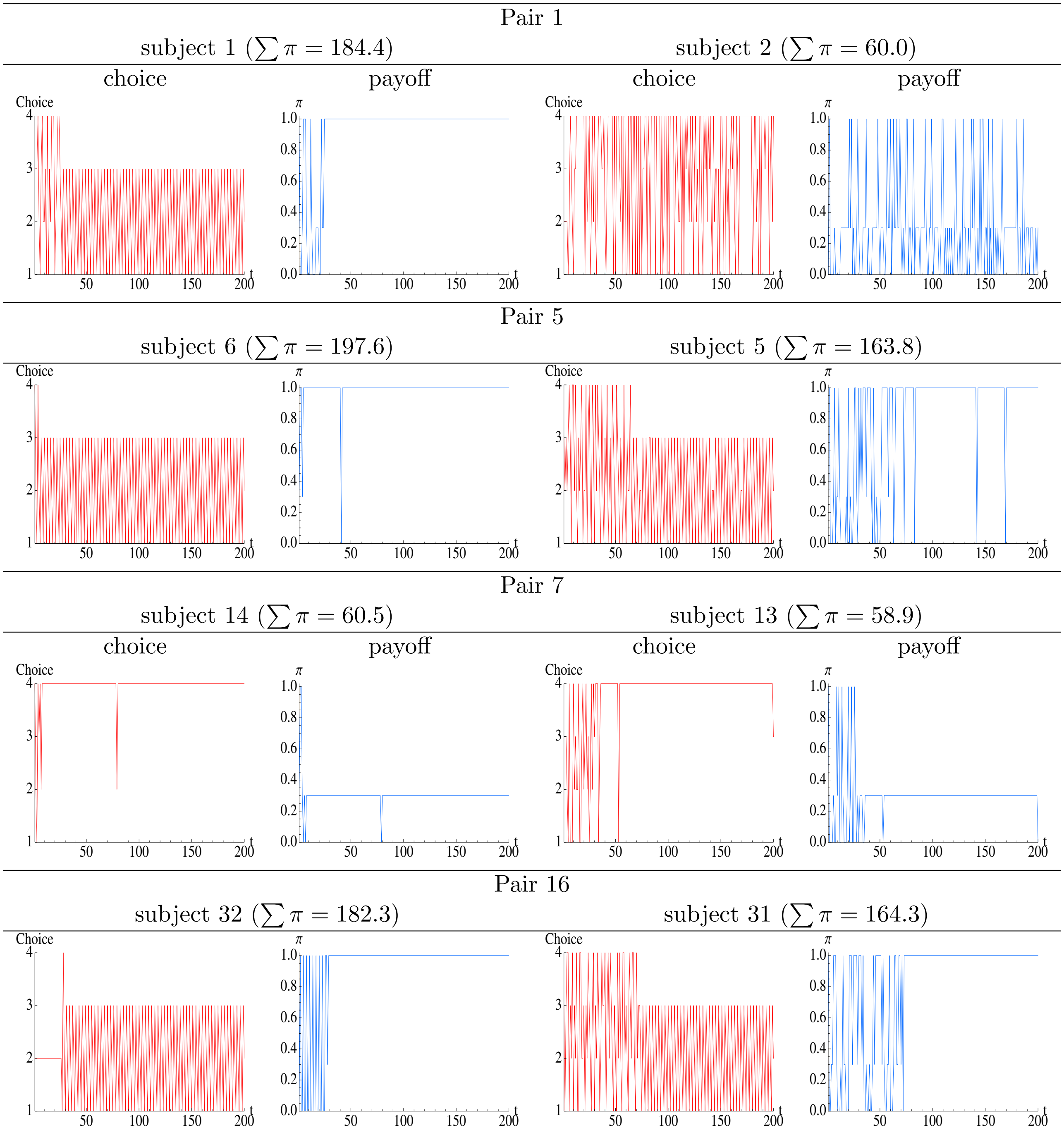
4.2 Extending the baseline model

In what follows we extend our baseline model to account for the additional information provided in each of the above treatments. We note, however, that subjects were not informed that if they chose the same arm as the other subjects with whom they are matched, they would receive the same pay-off. This reduces the inferences they could make from their observations.

4.2.1 The CPI case

We assume that agents keep track not only of their own choice and payoff, but also of the other's, and thus learn on the basis of these two information sets.[12]

[12] Bayer and Wu (2016) report evidence that suggests that subjects actually learn mostly from their own experience rather than from other's. However, their finding pertains to Bertrand duopolies which involve strategic interactions whereas our setting involves no such interactions.



Pair 1

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

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

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 ($\sum \pi = 58.9$)

Pair 16

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

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

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

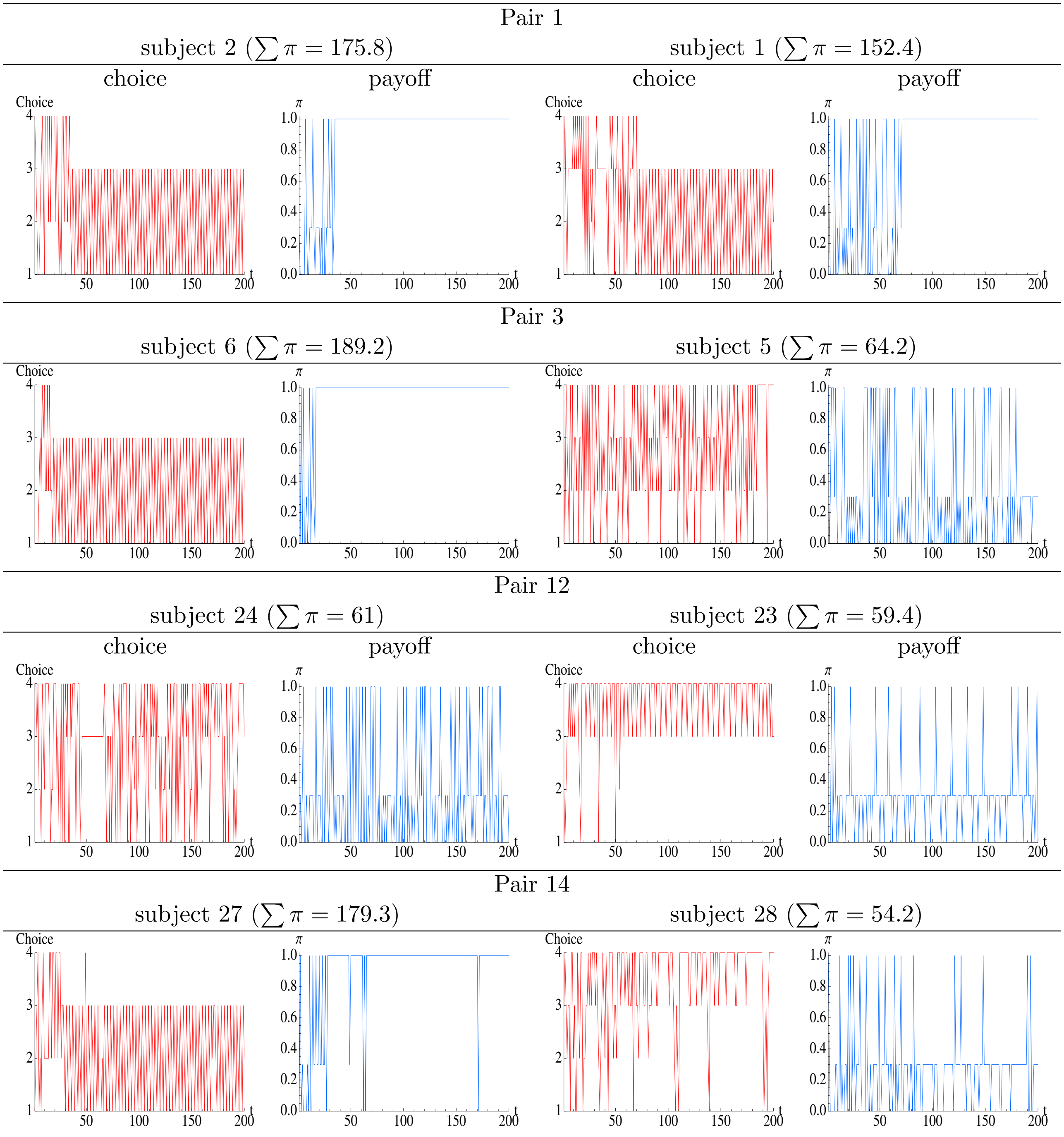


Figure 6: Examples of choice and payoff patterns in the PI treatment.

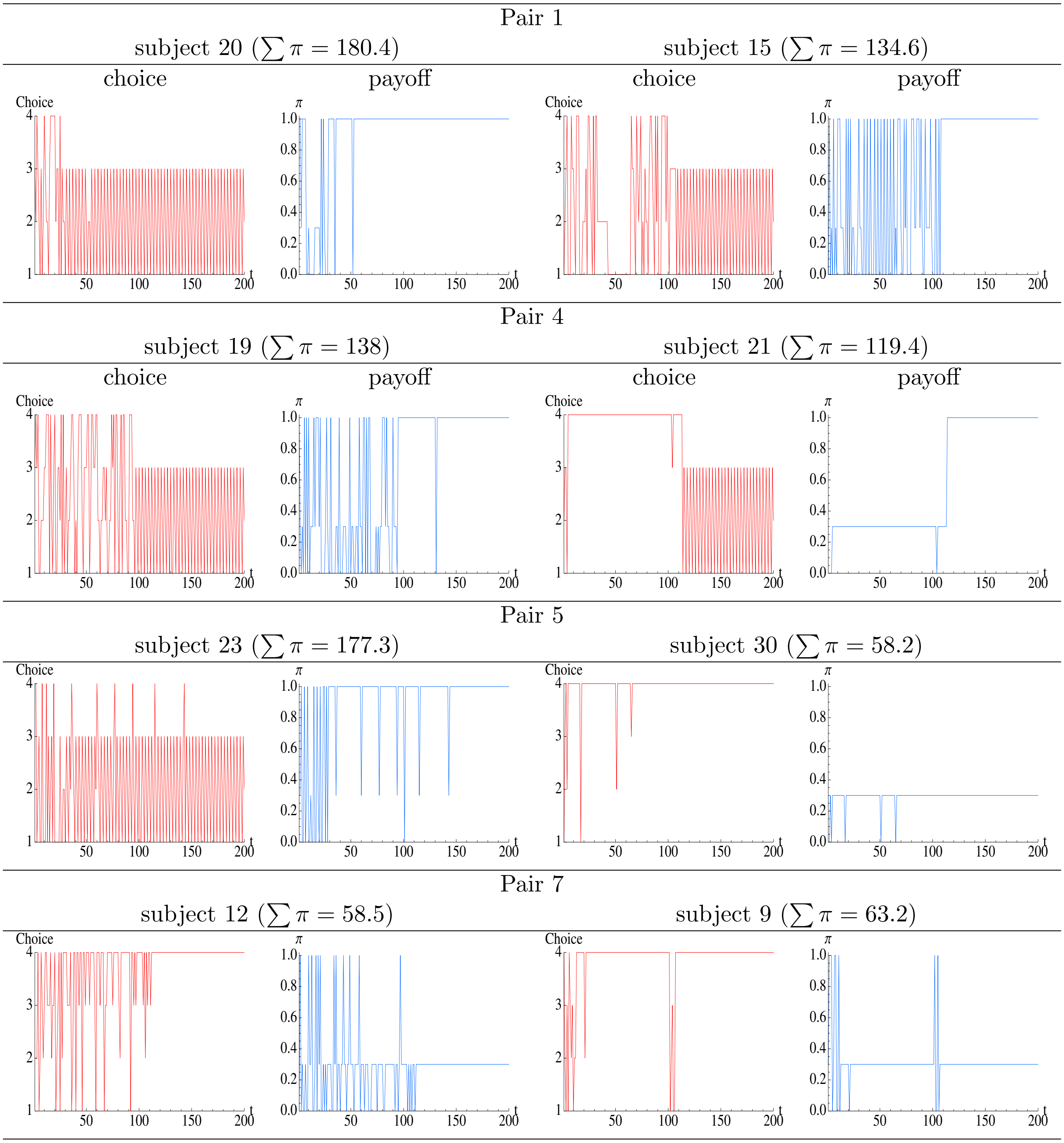
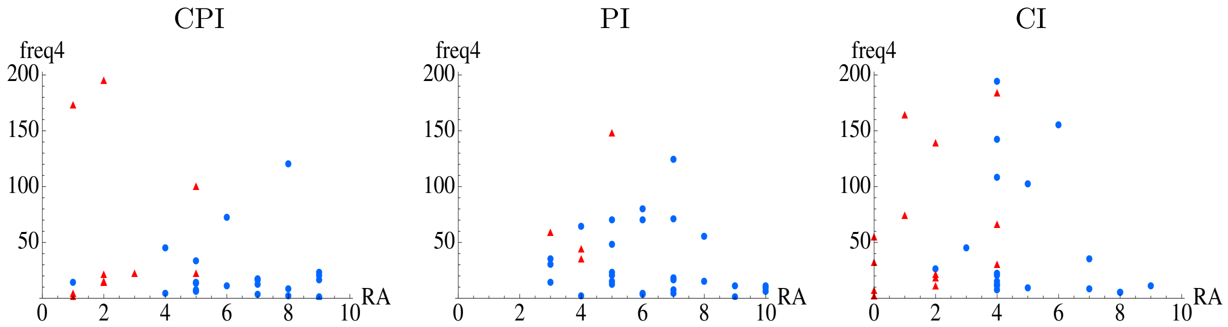


Figure 7: Examples of choice and payoff patterns in the CI treatment.



Without triangles

$N = 22$ $\quad$ $N = 28$ $\quad$ $N = 19$

$\rho = -0.094$ $(p = 0.680)$ $\quad$ $\rho = -0.282$ $(p = 0.148)$ $\quad$ $\rho = -0.413$ $(p = 0.079)$

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

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

Figure 8: Risk Aversion (x-axis) *vs* Frequency of Safe Option (y-axis). Triangle: subjects who have switched multiple times in Holt and Laury (2002) questionnaire.

Let $A_t^i(a)$ summarize, at the beginning of period 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:

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

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 on the sequence of his own (choice, payoff) pairs as well as on the sequence of j's (choice, payoff) pairs.

Let $B_t^i(a|h) = A_t^i(a)$ for all h from the moment when i becomes aware of the inter-temporal dependencies. We assume $B_t^i(a|h)$ to evolve as follows:

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

Based on $B_{t+1}^i(a|h)$, the choice of option a for agent i in period t is thus be defined as:

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

4.2.2 The 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 *(i)* what i knows about the set of possible (choice, payoff) pairs and *(ii)* their inter-temporal dependencies. On the one hand, if i is not yet aware of conditional strategies, then we assume that by observing j's sequence of high payoffs, i becomes aware of the existence of such strategies. Adopting the same reasoning, 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.

If i is already aware of conditional strategies and has just experienced the (choice, payoff) sequence $(1,1),(1,0)$, then i will infer (assuming he already knows that $\Pi(a)_t^i = \{0,1\}$ for $a \in \{1,2,3\}$ and $\Pi(4)_t^i = \{0.3\}$) 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 period and 0 in the current one.[13] 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 $A_t^i(a)$ evolves as follows:

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

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

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,

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

Again, we are basically assuming that j's payoff information is used only if j obtains a payoff of 1 in two consecutive periods. The choice of a conditional strategy in period $t+1$ will be based on $B^i_{t+1}(a|h)$ just as in the baseline case.

4.2.3 The CI case

Here, intuitively, the decision will involve mimicking. We will, however, implement it within our modeling framework through the evolution of attractions for consistency with the analysis of the NI treatment (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, i will assume that j is doing it because it generates high payoff. So 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 choices).

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 r^i times in the row. Say i has observed j making the sequence of choices $3, 2$ twice in row. If i has already learned about the performance of conditional strategies, this will translate into i assuming that strategy $2|(3,1)$ will result in payoff of 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 $A_t^i(a)$; and that this will happen only when i's choices are conditional.

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

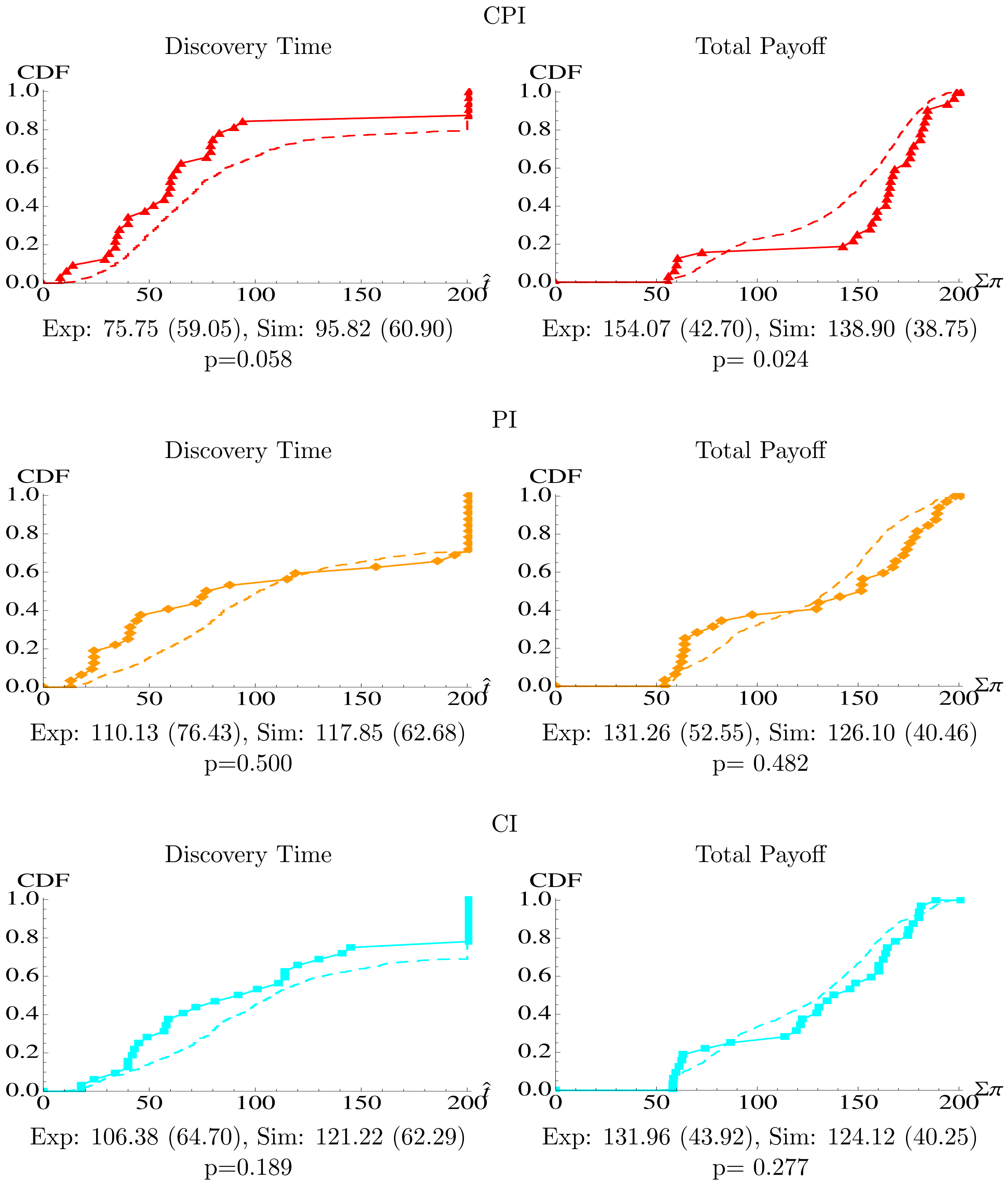
The choice of a conditional strategy in period $t+1$ will be based on $B_{t+1}^i(a|h)$ just as in the baseline case.

4.3 Simulation results

We simulate the above models using the parameter values best that best fit the baseline model described in Section 3. So for each individual in each pair of 500 simulated pairs of agents, we randomly draw (with replacement) each of the parameter values from the set $\{(\hat{\alpha}^i, \hat{\lambda}^i, \hat{r}^i)\}$ obtained for the NI treatment. The resulting distributions of individual discovery times and total payoffs are displayed in Figure 9 along with the observed ones.

The plots suggest that, overall, subjects discovered the hidden cycle somewhat faster than our simulated agents (c.f. left panels) which consequently led them to earn higher payoffs (c.f., right panels). However, this only partially holds for the CPI treatment since the outcomes of Permutation tests indicate a significant difference at $\alpha = 0.05$ between simulated and observed only for total payoffs ($p = 0.058$ for discovery times and $p = 0.024$ for total payoffs). Otherwise, we find no significant difference so that, as the reported average statistics also indicate, our modeling approach provides a remarkably good overall out-of-sample fit to the data ($p = 0.500$ in PI and $p = 0.189$ in CI for discovery times, and $p = 0.482$ in PI and $p = 0.277$ in CI for total payoffs).

We proceed with cross-treatment comparisons of the models' simulations and the observed data. The upper panel of Figure 10 displays the simulated distributions of discovery times (left panel) and total payoffs (right panel) whereas the lower panel displays the observed distributions along with the



CPI

Discovery Time

Exp: 75.75 (59.05), Sim: 95.82 (60.90)
p=0.058

Total Payoff

Exp: 154.07 (42.70), Sim: 138.90 (38.75)
p= 0.024

PI

Discovery Time

Exp: 110.13 (76.43), Sim: 117.85 (62.68)
p=0.500

Total Payoff

Exp: 131.26 (52.55), Sim: 126.10 (40.46)
p= 0.482

CI

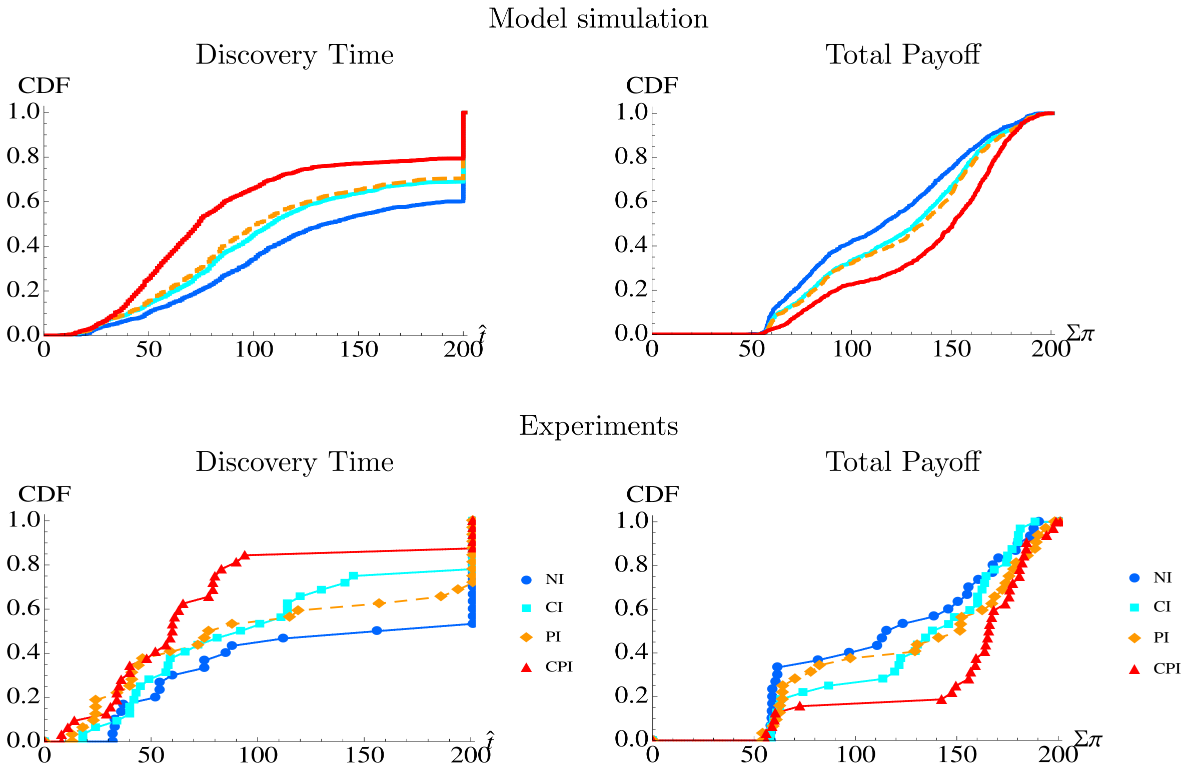
Discovery Time

Exp: 106.38 (64.70), Sim: 121.22 (62.29)
p=0.189

Total Payoff

Exp: 131.96 (43.92), Sim: 124.12 (40.25)
p= 0.277

Figure 9: CDFs of discovery times and total payoffs.
Legend: Observed (solid), Simulations (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

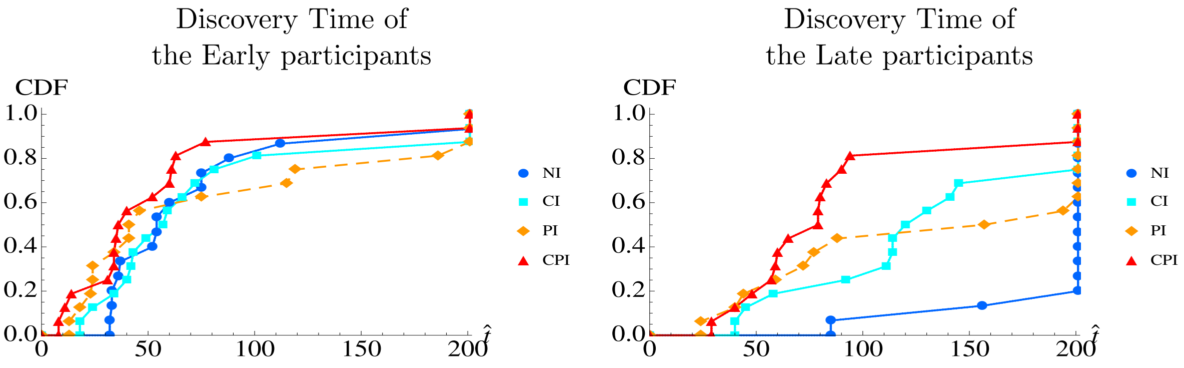
Figure 10: CDFs of discovery times and total payoffs (with p-values of Kruskal-Wallis tests for multiple sample comparisons and permutation tests for pair-wise comparisons (one-tailed)). Legend: NI (blue), CPI (red), PI (dashed orange), CI (light blue).

outcomes of Kruskal-Wallis (KW) and Permutation tests that assess their stochastic equivalence. Both the simulated and the observed distributions (and the test statistics) clearly support the conjecture that the more complete the information one has about the other subject, the faster the hidden pattern will be discovered. However, partial information of the type provided in our experiments participants receive (i.e, choices or payoffs) does not significantly affect the variables of interest; as both the simulated and the observed distributions indicate. The data actually suggest no significant difference between NI and CI or PI ($p > 0.05$) and significantly smaller (higher) discovery times (total payoffs) in CPI than in the other treatments.

4.4 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 will help the observer to discover the hidden 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 subjects, the one who discovered the hidden pattern first, henceforth the Early participant, or second, henceforth the Late participant. In case the hidden pattern was not discovered, the labels are randomly assigned. We also randomly match participants in the NI treatment in pairs and assign labels by comparing their respective discovery times to get some benchmark for Early and Late participants who do not observe any other participant.

Figure 11 shows the distributions of discovery times of each type of participant in each treatments along with the outcomes of nonparametric tests to assess the observed differences. The plots and test statistics for Late participants (right panel) indicate that they greatly benefited from this additional information. It appears that those in CPI benefited most and significantly more than those in CI or PI whose distributions are not significantly different. The plots also indicate that when compared to the NI treatment, the partial information feedback of these treatments significantly improved their time of discovery. However, for the Early participants (left panel) we find no significant difference across treatments so that the additional information provided in the information treatments had no significant effect on the performance of Early participants.



$p = 0.527$ (KW)

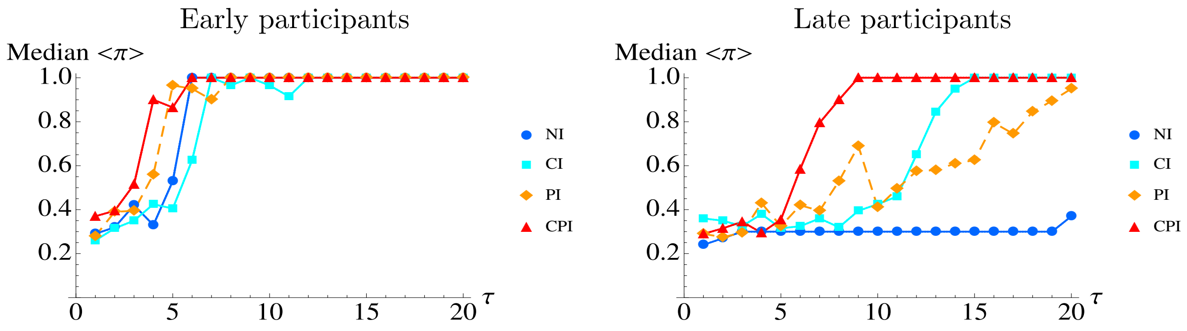
p-values	CI	PI	CPI
NI	0.575	0.647	0.221
CI		0.575	0.172
PI			0.147

$p < 0.001$ (KW)

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

Figure 11: CDFs of discovery times of Early and Late participants (with p-values of Kruskal-Wallis tests for multiple sample comparisons and permutation tests for pair-wise comparisons (one-tailed)). Legend: NI (blue), CPI (red), PI (dashed orange), CI (light blue).

Next, to determine who benefited most from the information provided, we tracked the evolution of average payoffs of each type in the four treatments. Figure 12 reports the median of the average pay-offs from each round for Early and Late participants for each block of 10 rounds in each treatment. The plots and the reported KW test statistics (c.f. lower panel of Figure 12) suggest that round average payoffs are not significantly different across treatments for the first five blocks and are actually very similar across types. Differences emerge from the sixth block onwards as the round average payoffs of Late participants vary widely across treatments whereas those of Early participants become identical. This confirms that the information provided did not affect the discovery time of Early participants whereas it significantly helped Late participants in the three information treatments considered. To repeat, observing a successful subject is very helpful but observing an unsuccessful subject does not speed up learning.



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 12: Dynamics of median round average payoff (over block of 10 rounds) of Early and Late participants (with *p*-values of Kruskal-Wallis tests).
Legend: NI (blue), CI (solid light blue), PI (dashed light blue), CPI (red).

5 Concluding remarks

This paper examines the problem of how people learn to improve the results that they obtain from taking different actions. In particular we focus on situations in which individuals have no knowledge of the structure of the environment within which they operate. In many problems of this sort, it is assumed that there is a simple mapping from actions to pay-offs. We looked at a situation in which pay-offs depend on the sequence of actions taken. To make improvements in such a case, subjects, who may start by reinforcing on the pay-offs from single actions, have to realise that there is an inter-temporal dependence.

Our experiments involved multi-armed bandits with a hidden deterministic payoff structure and we wished to see whether subjects could discover this structure or to behave as if they had done so. In a baseline treatment (NI), subjects were only informed of the payoff they themselves have obtained after each choice. While some of our subjects were surprisingly fast in “discovering” the hidden mechanism in this treatment, half of them failed to do so within 200 rounds. Most of our subjects initially explored among four different options. At some point, however, those who were successful seem to had, what has been referred to in the psychology literature as an "aha" moment, and started to reinforce on sequences of actions taken in the past. We developed a simple model that tries to incorporate this type of "counter intuitive learning" and found that it generates (simulated) distributions of discovery times and of total payoffs that match remarkably well those observed.

We then considered a further series of experiments through which we examined how the possibility of observing what others do or obtain may help subjects to improve their gains. This is the problem that is raised in the literature on "observational learning." In these additional paired treatments, a subject could, in addition to the own payoff, observe the last choice made (CI), the last payoff obtained (PI), or both (CPI) by another subject. In these treatments, however, the additional observation did not seem to significantly facilitate the task for subjects who were the first in their pair to discover the pattern to do so compared to those in the NI treatment who discovered the pattern earlier in randomly created pairs. However, compared to those who were slower to discover the pattern in NI treatment, those subjects in paired treatments who discovered the pattern later than the other discovered it significantly faster. This was true even if subjects were only observing the payoffs, and not the choices, obtained by the other and thus had to discover the exact choice sequence leading to consecutive payoffs of one, the maximum attainable in our framework. As might be expected, the CPI case led to the most improvement for these slower subjects.

We extended our basic model to capture the observed behaviour in these additional experiments. We found that our model gave discovery times which corresponded well to the experimental evidence in the CI and the PI treatments. In the CPI treatment, however, our model resulted in later discovery times than was observed in our experiment. This gap between the model and the experimental observations may reflect some compounding effect of information that is not captured well in our model.

We believe our modelling approach is the first to study this type of learning with(out) observational learning and hope that it will induce further research that will investigate whether our results are specific to the particular deterministic payoff generating mechanism we employed in our experiment. For example, one can investigate how the existence of a safe option that returned a low but sure payoff would affect the results by eliminating such an option or changing it to a risky option with the same expected payoff. One could also extend the memory of the subjects, by displaying the past outcomes, to see if this facilitates their learning.

Another fruitful avenue for future research would be to consider several models of learning that allows an agent to become aware of features of the environment and start expanding the set of possible choice strategies, and compare which model does a better job in replicating the observed experimental outcomes by running a competition *à la* Erev et al. (2016). Among many possibilities, building upon the models proposed by Donoso et al. (2014) or Plonsky et al. (2015) seem to be fruitful ways forward.

Finally, as we have mentioned, a number of the questions that are raised in this paper are related to issues raised in the psychology and in the neuroscience literature. For example, to better understand the mechanisms involved in producing an "aha" moment, a moment in which the subjects change their perspective on the environment about which they are learning, would be of considerable interest. To understand the neural processes at work when such an insight occurs, possibly by using fMRI or other instruments, is a promising avenue for future research.

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.

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.

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.

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.

Castle, J. L., J. A. Doornik, and D. F. Hendry (2013): "Evaluating Automatic Model Selection," Tech. rep., University of Oxford.

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.

Donoso, M., A. G. E. Collins, and E. Koechlin (2014): "Foundations of human reasoning in the prefrontal cortex," *Science*, 344, 1481–1486.

Dufour, J.-M. (2008): "Model selection," in *The New Palgrave Dictionary of Economics*, ed. by S. N. Durlauf and L. E. Blume, Palgrave Macmillan, second edition ed.

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., E. Ert, O. Plonsky, D. Cohen, and O. Cohen (2016): "From anomalies to forecasts: Toward a descriptive model of decision under risk, under ambiguity, and from experience," mimeo, Technion.

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.

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

Greiner, B. (2015): "An Online Recruitment System for Economic Experiments," *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.

March, J. G. (1991): "Exploration and Exploitation in Organizational Learning," *Organization Science*, 2, 71–87.

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.

Plonsky, O., K. Teodorescu, and I. Erev (2015): "Reliance on Small Samples, the Wavy Recency Effect, and Similarity-Based Learning," *Psychological Review*, 122, 621–647.

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.

Stephens, D. W. and J. R. Krebs (1986): *Foraging theory*, Princeton, NJ: Princeton University Press.

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.

6 Appendix

The experiments were conducted in the laboratory for experimental economics of the Australian Business School (*ASBLab*) at the University of New South Wales. Participants were students in Economics, Commerce, Marketing, and Engineering who were recruited on campus using the ORSEE software (Greiner, 2015). Upon arrival in the laboratory, they were randomly assigned to individual cubicles and given the following sets of instructions that were read aloud. Once the experiments finished, they were individually called to collect their reward in cash. The experiments were programmed using *z-tree* (Fischbacher, 2007).

6.1 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

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

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.