Rage against the machines: how subjects play against learning algorithms

Peter Duersch · Albert Kolb · Jörg Oechssler · Burkhard C. Schipper

Received: 15 February 2008 / Accepted: 26 January 2009 / Published online: 19 February 2009
© The Author(s) 2009. This article is published with open access at Springerlink.com

Abstract We use a large-scale internet experiment to explore how subjects learn to play against computers that are programmed to follow one of a number of standard learning algorithms. The learning theories are (unbeknown to subjects) a best response process, fictitious play, imitation, reinforcement learning, and a trial & error process. We explore how subjects’ performances depend on their opponents’ learning algorithm. Furthermore, we test whether subjects try to influence those algorithms to their advantage in a forward-looking way (strategic teaching). We find that strategic teaching occurs frequently and that all learning algorithms are subject to exploitation with the notable exception of imitation.

Keywords Learning · Fictitious play · Imitation · Reinforcement · Trial & error · Strategic teaching · Cournot duopoly · Experiments · Internet

JEL Classification C72 · C91 · C92 · D43 · L13

Financial support by the DFG through SFB/TR 15 and SFB 504 is gratefully acknowledged. We thank two anonymous referees, David Cooper, Drew Fudenberg, Tim Grebe, Aaron Lowen, and seminar participants in Edinburgh, Heidelberg, Mannheim, Vienna, Tsukuba, the University of Arizona, and at the ESA Meetings 2005 in Tucson for helpful comments.

P. Duersch · A. Kolb · J. Oechssler (✉)
Department of Economics, University of Heidelberg,
Grabengasse 14, 69117 Heidelberg, Germany
e-mail: oechssler@uni-hd.de

P. Duersch
e-mail: peter.duersch@awi.uni-heidelberg.de

B. C. Schipper
Department of Economics, University of California, Davis,
One Shields Avenue, Davis, CA 95616, USA
e-mail: bcschipper@ucdavis.edu
1 Introduction

In recent years, theories of learning in games have been extensively studied in experiments. The focus of those experiments was primarily the question which learning theories describe best the average behavior of subjects. It turns out that some very simply adaptive procedures like reinforcement learning, best response dynamics, or imitation are fairly successful in describing average learning behavior of subjects in some games (see, e.g., Erev and Haruvy 2008, for a recent survey).

The focus of the current experiment is different. We are interested in the following strategic aspects of learning in games. First, how is a player’s success affected by the way opponents learn? Second, how can the opponent’s learning process be influenced by the player’s behavior? For example, can it be manipulated to the player’s advantage? To address those questions, we present here a first—very exploratory—experimental study. Since we are interested in how subjects respond to certain learning theories, we need to be able to control the behavior of opponents. The best way to do that is by letting subjects play against computers programmed with particular learning theories. 1

The questions raised in this paper seem to be fairly novel, although the second question has received some attention at least in the theoretical literature. 2 For example, Fudenberg and Levine (1998, p. 261) write “A player may attempt to manipulate his opponent’s learning process and try to ‘teach’ him how to play the game. This issue has been studied extensively in models of ‘reputation effects’, which typically assume Nash equilibrium but not in the context of learning theory.” Following Camerer and Ho (2001) and Camerer et al. (2002) we shall call this aspect of learning “strategic teaching”. 3 We believe that this hitherto largely neglected aspect of learning is of immense importance and deserves further study. As we shall see in this experiment, theories just based on adaptive processes will not do justice to the manipulation attempts of subjects.

We consider five learning theories in a Cournot duopoly: best-response (br), fictitious play (fic), imitate-the-best (imi), reinforcement learning (re), and trial & error (t&e). Some noise is added in order to make the task less obvious. Noise is also a requirement for some of the theoretical predictions to work as it prevents a learning process from getting stuck at states which are not stochastically stable. 4 The selection of learning theories is based on the prominence in the literature, convenient applicability to the Cournot duopoly, and sufficient variety of theoretical predictions.

The experiment was conducted as a large scale internet experiment. Internet experiments are still relatively novel (see, e.g., Drehmann et al. 2005, for first experiences). Arguably, the setting (working at home or in the office at your own PC) is more

---

1 Subjects are, of course, being told that they play against computers. There is now a large experimental literature making use of computers to control for some players’ behavior in strategic situations, see Cason and Sharma (2007) for a recent experiment.

2 See Fudenberg and Levine (1989) and Ellison (1997).

3 Note, however, that we use the term in a broader sense, not necessarily referring to EWA as in Camerer et al. (2002).

4 See, e.g., Vega-Redondo (1997) for imitate-the-best and Huck et al. (2004a) for trial & error.
representative of real world decisions than in the usual laboratory experiments. Also, internet experiments allow to reach a large subject pool at moderate cost.\(^5\)

With respect to the first question, we find that subjects achieve substantially higher profits than all of their computer opponents but one. The exception is the imitation algorithm, for which we show theoretically that it cannot be beaten by more than a small margin and which in fact performs on average better than its human opponents in the experiment. The computer opponent that allows for the highest profits for its human counterparts is the reinforcement learning computer. However, due to the stochastic nature of reinforcement learning, a lot of luck is needed, and the variances are high.

This leads us to the second question: We find that strategic teaching occurs frequently and that all learning algorithms are subject to exploitation with the notable exception of imitation. Subjects learn quickly how to exploit the best response—and trial & error—computers, usually by behaving as Stackelberg leader, although some subjects manage to find more innovative and even more profitable ways.

Two papers are closely related to our work. Shachat and Swarthout (2002) let subjects play against both human subjects and computers, which are programmed to follow reinforcement learning or experienced weighted attraction in repeated 2 × 2 games with a unique Nash equilibrium in mixed strategies. They find that human play does not significantly vary depending on whether the opponent is a human or a programmed learning algorithm. In contrast, the learning algorithms respond systematically to non-Nash behavior of human subjects. Nevertheless, these adjustments are too small to result in significant payoff gains. Coricelli (2005), on the other hand, found that human subjects do manage to exploit computer opponents that play a biased version of fictitious play in repeated 2 × 2 zero-sum games.

The remainder of the paper is organized as follows. Section 2 describes the Cournot game that is the basis for all treatments. In Sect. 3 we introduce the computer types and the associated learning theories. The experimental design is explained in Sect. 4, followed by the results in Sect. 5. In Sect. 6 we consider a laboratory treatment as a robustness check. Section 7 concludes. The instructions for the experiment and screenshots are shown in the Appendix.

2 The Cournot game

We consider a standard symmetric Cournot duopoly with linear inverse demand function \(\max\{109 - Q, 0\}\) and constant marginal cost, \(MC = 1\). Each player’s quantity \(q_i, i = 1, 2\) is an element of the discrete set of actions \(\{0, 1, \ldots, 109, 110\}\). Player \(i\)’s profit function is given by

\[
\pi(q_i, q_{-i}) := \left(\max\{109 - q_i - q_{-i}, 0\} - 1\right) q_i.
\]

Table 1 shows outputs and profits for the Nash equilibrium, the competitive outcome (where \(p = MC = 1\)), the collusive outcome, the Stackelberg outcome, and

\(^5\) Since internet experiments are relatively novel, we explore some methodological issues of this experiment in a companion paper by comparing it to various laboratory treatments (see Duersch et al. 2008b).
Table 1 Prominent outcomes

| Outcome                        | $q_i$ | $q_{-i}$ | $\pi_i$ | $\pi_{-i}$ |
|-------------------------------|-------|----------|---------|------------|
| Cournot-Nash equilibrium      | 36    | 36       | 1,296   | 1,296      |
| Symmetric competitive outcome | 54    | 54       | 0       | 0          |
| Symmetric collusive outcome   | 27    | 27       | 1,458   | 1,458      |
| Stackelberg leader outcome    | 54    | 27       | 1,458   | 729        |
| Stackelberg follower outcome  | 27    | 54       | 729     | 1,458      |
| Monopoly solution             | 54    | 0        | 2,916   | 0          |

the monopoly solution. Subjects play the Cournot duopoly repeatedly for 40 rounds. Thus, we index the quantity $q_t^i$ by the period $t = 1, \ldots, 40$.

A Cournot duopoly is chosen for this experiment because, based on earlier theoretical and experimental contributions, we expected that the behavior of the various learning theories would differ in interesting ways in a Cournot game. In particular, there was the conjecture that imitation would behave very differently from the remaining learning theories. In order to make this conjecture precise, we derive in this paper a new theoretical result, namely that the imitation algorithm cannot be beaten by much even by a very sophisticated player. Of course, this result applies only to a particular class of games that includes the Cournot game but also games like chicken.6

3 Computer types

Computers were programmed to play according to one of the following decision rules: Best-response (br), fictitious play (fic), imitate the best (imi), reinforcement learning (re), or trial & error (t&e). All decision rules except reinforcement learning are deterministic, which would make it too easy for subjects to guess the algorithm (as we experienced in a pilot study to this project). Therefore, we introduced some amount of noise for the deterministic processes (see below for details). The action space for all computer types was $\{0, 1, \ldots, 109\}$.

All computer types require an exogenously set choice for the first round as they can only condition on past behavior of subjects. To be able to test whether starting values matter, we chose different starting values. However, to have enough comparable data, we restricted the starting values to 35, 40, and 45. Starting quantities were switched automatically every 50 plays in order to collect approximately the same number of observations for each starting quantity but subjects were unaware of this rule.

3.1 Best-response (br)

Cournot (1838) himself suggested a myopic adjustment process based on the individual best-response

6 We thank a referee for this observation.
\[ q_i^t = \arg \max_{q_i} \pi(q_i, q_{-i}^{t-1}) = \max \left\{ \frac{108 - q_{-i}^{t-1}}{2}, 0 \right\}, \] (2)

for \( t = 2, \ldots \). Moreover, the parameters are such that if both players use the best-response process, the process converges to the Nash equilibrium in a finite number of steps (see, e.g., Monderer and Shapley 1996).

This deterministic process is supplemented by noise in the following way. If the best response process yields some quantity \( q_i^t \), the computer actually plays a quantity chosen from a Normal distribution with mean \( q_i^t \) and standard deviation 2, rounded to the next integer in \( \{0, 1, \ldots, 109\} \).\(^7\) This implementation of noise is also used for computer types fictitious play and imitation.

3.2 Fictitious play (fic)

A second decision rule that is studied extensively in the literature is fictitious play (see Brown 1951; Robinson 1951; Fudenberg and Levine 1998, Chap. 2). A “player” (that is, in our setting, the computer) who uses fictitious play chooses in each round a myopic best response against the historical frequency of his opponent’s actions (amended by an initial weight for each action). If we let those initial weights be the same for each action and each player, \( w_1^i(q_{-i}) = w^1 \), we obtain the following recursive formulation for the weight player \( i \) attaches to his opponent’s action \( q_{-i} \), where 1 is added each time the opponent chooses \( q_{-i} \).

\[ w_i^t(q_{-i}) = w_i^{t-1}(q_{-i}) + \begin{cases} 1 & \text{if } q_{-i}^{t-1} = q_{-i} \\ 0 & \text{if } q_{-i}^{t-1} \neq q_{-i} \end{cases} \]

for \( t = 2, \ldots \). Player \( i \) assigns probability

\[ p_i^t(q_{-i}) = \frac{w_i^t(q_{-i})}{\sum_{q'_{-i}} w_i^t(q'_{-i})} \]

to player \(-i\) using \( q_{-i} \) in period \( t \). Consequently, player \( i \) chooses a quantity that maximizes his expected payoff given the probability assessment over the opponent’s quantities, i.e.,

\[ q_i^t \in \arg \max_{q_i} \sum_{q_{-i}} p_i^t(q_{-i}) \pi(q_i, q_{-i}). \] (3)

We simulated the fictitious play processes against itself and some other decision rules for many different initial weights \( w^1 \) and ended up choosing \( w^1 = 1/25 \). Except for much smaller or much larger initial weights, results of the simulations did not

\( ^7 \) Due to a programming error in the rounding procedure, the noise of computer types br, fic, and imi was actually slightly biased downwards (by 0.5), which makes the computer player slightly less aggressive. This does not have any lasting effects for computer types br and fic but has an effect on imi.
change much. Very high initial weights lead to rather slow adaptation whereas very small ones resulted in erratic movements. Since our Cournot duopoly is a potential game, fictitious play converges to the unique Cournot-Nash equilibrium (see Monderer and Shapley 1996).

3.3 Imitate the best (imi)

Imitation has received much attention recently in both theory and experiments (see e.g., Vega-Redondo 1997; Apesteguia et al. 2007; Schipper 2008). The rule “imitate the best” simply requires to choose the best action that was observed in the previous period. If player $i$ follows this decision rule in $t = 2, \ldots$, he chooses

$$q_t^i = \begin{cases} q_t^{i-1} & \text{if } \pi(q_t^{i-1}, q_{-i}^{t-1}) \geq \pi(q_{-i}^{t-1}, q_t^{i-1}) \\ q_{-i}^{t-1} & \text{otherwise.} \end{cases}$$

(4)

Vega-Redondo (1997) shows for a symmetric Cournot oligopoly that if all players follow this decision rule up to a small amount of noise, then the long run distribution over quantities assigns probability 1 to the competitive outcome (where $p = MC$) as the noise vanishes.\(^8\) The intuition for this is simple. For all total quantities $Q$ such that $p > 1$, the firm with the highest quantity receives the highest profit. When the highest quantity is imitated by other firms, $Q$ increases. For $Q$ such that $p < 1$, the firm with the lowest quantity is being imitated, such that $Q$ decreases. Thus, $Q$ converges to the competitive quantity. See Huck et al. (1999) and Offerman et al. (2002) for experimental evidence.

With respect to the current experiment, of particular interest is the question what happens when only one player imitates the best? The following proposition shows that imi can essentially not be exploited even by very sophisticated players.\(^9\)

**Proposition 1** Suppose a player knows that his opponent follows the rule imitate the best (imi) and even knows the opponent’s starting value.

(a) If this player wants to maximize his absolute payoff over 40 rounds, then the optimal strategy yields an average payoff of 46,374, which is much less than the profit of 55,068 for his computer opponent imi and also less than the Cournot profit of 51,840.

(b) If the player wants to maximize his relative payoff (i.e., the difference between his payoff and his opponent’s payoff), then an optimal strategy yields an average profit differential of 212.67.

---

\(^8\) Vega-Redondo’s result has been generalized to larger classes of games by Alós-Ferrer and Ania (2005), Possajennikov (2003), Schipper (2003) and Tanaka (1999).

\(^9\) Schipper (2008) shows that if there are both imitators and best-response players in the game, then any state where imitators are weakly better off than best-response players and where best-response players play a best-response is absorbing. Moreover, if mistakes are added, then in the long run imitators are strictly better off than best-response players. The intuition is that if imitators play a sufficiently large quantity, best–responders become Stackelberg followers.
The implications of the proposition are the following. First, even sophisticated human players that maximize their long-run profits in a forward looking way against imi will achieve much lower payoffs than their computer opponent imi. Second, the total profit of the sophisticated player is much lower than the total profit of the stage game Cournot-Nash equilibrium over 40 rounds or the profit a Stackelberg leader could achieve against computer br. Hence, even sophisticated human subjects playing against imi will typically earn much less than playing against other computer types. Third, since imi will never lower its quantity \( q_2 \) as long as \( q_1 + q_2 < 108 \), any mistake or experimentation that leads to an increase of the human subject’s quantity will result in a permanent drop of profits. Thus, we should expect to see profits decline over time. Finally, even subjects that do not care to maximize their absolute profits but instead aim at beating the computer (i.e., maximize their relative payoff), can only do so by a very modest margin of 212 (for comparison note that a Stackelberg leader gains a profit differential against a follower of \( 40 \times 729 = 29,160 \)).

3.4 Reinforcement learning (re)

In a standard model of reinforcement learning by Roth and Erev (1995), an action is chosen with probability that is proportional to the propensity for this action. Propensities, in turn, are simply the accumulated payoffs from taking this action earlier in the process.

In games with a large action space such as a Cournot duopoly, it seems unreasonable to reinforce only that single action that was chosen in a given round. Rather, actions in the neighborhood should also be reinforced although to a lesser extent depending on their distance to the original choice. Therefore, we complement the standard model of reinforcement learning by updating of neighborhoods à la Sarin and Vahid (2004).

The player starts with the same initial propensity for each quantity, \( w_{i}^{1}(q) \). For \( t = 2, \ldots \), propensities are updated by\(^{10}\)

\[
w_{i}^{t}(q) = \max \left[ 1, w_{i}^{t-1}(q) + \beta(q, q_{i}^{t-1})\pi(q_{i}^{t-1}, q_{-i}^{t-1}) \right],
\]

where \( \beta \) is the linear Bartlett function

\[
\beta(q, q_{i}^{t-1}) := \max \left\{ 0, \frac{6 - |q - q_{i}^{t-1}|}{6} \right\}.
\]

That is, all actions within five grid points of the chosen action are also reinforced.

\(^{10}\) We imposed a lower bound of 1 on propensities.
The probability of playing quantity \( q \) in period \( t \) is computed by normalizing the propensities

\[
p'_i(q) = \frac{w'_i(q)}{\sum_{q'} w'_i(q')}.
\]

Theoretical results on the convergence properties of reinforcement learning are scarce.\(^{11}\) Thus most of the analysis is based on simulations. We ran several simulations of reinforcement learning against itself as well as other decision rules while varying the initial propensities \( w'_i(q) \). Choosing initial propensity is always a bit arbitrary. However, results did not change much when using different initial propensities. We chose \( w'_i(q) = 78 \), which minimized the mean squared deviation to the Nash equilibrium. Since reinforcement learning already is a stochastic process, we did not add additional noise to the process.

3.5 Trial & error (t&e)

Huck et al. (2004a,b) introduce a very simple trial & error learning process. Players begin by randomly adjusting their initial quantity either up- or downwards with an exogenously fixed step size. If this change increases profits, the direction is continued. If it does not, the direction of adjustment is reversed. We chose a step size of 4. Formally, players adjust their quantities as follows:

\[
q'_i(t) := \max\{0, \min\{q'_i(t-1) + 4s'_i(t-1), 109\}\},
\]

for \( t = 2, \ldots \), where

\[
s'_i(t) := \begin{cases} 
\text{sign}(q'_i(t) - q'_i(t-1)) \times \text{sign}(\pi'_i(t) - \pi'_i(t-1)) & \text{if } (q'_i(t) - q'_i(t-1))(\pi'_i(t) - \pi'_i(t-1)) \neq 0 \\
+1, -1 & \text{each with positive probability}
\end{cases}.
\]

On the boundaries of the output grid, we chose a “soft reflecting boundary”. In particular, when the process reached 109 or 0 twice in subsequent periods, the next quantity chosen was 109 – 4 or 0 + 4, respectively.

Huck et al. (2004a) show that in Cournot duopoly if players choose the wrong direction with small but positive probability, then trial & error learning converges in the long run to a set of outcomes around the collusive outcome. To follow the theoretical setting, the noise for this process was modelled such that the computer chose the

\(^{11}\) Laslier et al. (2001) show that reinforcement learning converges with positive probability to any strict pure Nash equilibrium in finite two-player strategic games. Similar results were obtained by Ianni (2002). However, they do not consider reinforcement of neighborhoods as in our case.
opposite direction from that prescribed by the theory with independent probability of 0.2 in each round.  

4 Experimental design

In total, 550 subjects participated in our internet experiment. Subjects played in a location of their own choice (home, office etc.), and at their own pace. Recruitment was done by email, internet newsgroups, and a University of Bonn student magazine. Each subject chose a nickname. Incentives were provided by publicly displaying a highscore after the experiment (like in computer games). We did not want to exclude (and implicitly select) subjects by technical means. In order to participate in our experiment, a standard web browser and a low-speed internet connection were sufficient.

Subjects could repeat the experiment as often as they desired, either immediately or at some later time. Subjects were encouraged to repeat under the same user name as before. While 550 subject played the first round (“first-timers”), we recorded 500 plays by “repeaters”.

The sequence of events was as follows. After logging in, subjects were matched to a computer type. The computer type was displayed to subjects via a label (Greek letters) though subjects were not told how computer types were associated with labels. In the instructions (see Appendix A) subjects were told the following: “The other firm is always played by a computer program. The computer uses a fixed algorithm to calculate its output which may depend on a number of things but it cannot observe your output from the current round before making its decision.” A page with instructions was displayed to subjects. At any time during the experiment, subjects were able read the instructions and an example for calculating profits by opening a separate window on their computer. After reading the instructions, subjects could input their quantity for the first round. The computer displayed a new window with the results for the current round including the number of the round, the subject’s quantity, the subject’s profit, the computer’s quantity as well as the computer’s profit (see Appendix B for screenshots). Subjects had to acknowledge this information before moving on to the following round. Upon acknowledgment, a new page appeared with an input field for the new quantity. This page also showed a table with the entire history of previous rounds’ quantities and profits for both players.  

---

12 Trial & error learning can be viewed as a special operationalization of learning direction theory by Selten and Buchta (1998). This theory assumes that players can judge in which direction better actions can be found. In the absence of information about demand and cost conditions, one interpretation is that the right direction can be found by determining which direction was successful last period.

13 One subject actually played a record 31 times.

14 The incentives for doing so were the highscore and the possibility to pick the same computer opponent as before (subjects logging in under a different name were allocated to a randomly chosen computer). The latter possibility was only revealed once subjects logged in under the same name.

15 See the working paper version (Duersch et al. 2008a) for an additional treatment, in which subjects were reminded only of the previous round’s results. The results did not differ significantly.
Table 2  Mean quantities

| Treatment | Subjects’ Mean quantities | Computers’ Mean quantities |
|-----------|---------------------------|---------------------------|
| br        | 51.99 (0.61)              | 27.79 (0.30)              |
| t&e       | 48.96 (0.71)              | 32.05 (0.49)              |
| fic       | 46.11 (0.74)              | 31.94 (0.26)              |
| imi       | 46.40 (0.91)              | 48.38 (0.49)              |
| re        | 47.45 (0.83)              | 35.71 (0.72)              |
| Total     | 48.68 (0.34)              | 33.92 (0.29)              |

Note: Average quantities over all 40 rounds and all subjects in a given treatment. The Cournot-Nash equilibrium quantity is 36. Standard errors of means in parentheses.

After round 40, subjects were asked to fill in a brief questionnaire (see Appendix) with information on gender, occupation, country of origin, formal training in game theory or economic theory, previous participation in online experiments, and the free format question “Please explain in a few words how you made your decisions”. It was possible to skip this questionnaire. The highscore was displayed on the following page. This table contained a ranking among all previous subjects, separately for subjects who were matched against the same computer type and for all subjects. It also contained the computer’s highscore.

5 Results

To give a first impression of the data, we present in Table 2 mean quantities of subjects and computers, respectively, averaged over all rounds and subjects. The first thing to notice is that subjects on average have much higher quantities than computers (48.68 vs. 33.29). This holds for all treatments except for the imitation treatment. Recall that the Cournot-Nash quantity is 36 (see Table 1). Thus, subjects chose on average quantities that exceeded by far the Cournot quantity and in some cases came close to the Stackelberg leader output of 54.

5.1 How are profits affected by the opponent’s learning algorithm?

How do subjects’ profits differ with respect to their computer opponents? Figure 1 shows a boxplot, which compactly summarizes the range of subjects’ average profits per round of first time players and repeaters, respectively. The figure reports those measures separately for each treatment, i.e., for each computer opponent (br, t&e, fic, imi, and re). In the boxplot, the boxes denote the interquartile range (ICR) between 75th and 25th percentiles, i.e., 50% of observations are concentrated in this range. The line in the box denotes the median profit. The length of the whiskers is the min of 1.5 times the ICR and the distance to the most extreme outcome. The dotted line shows the profit per round in the Cournot-Nash equilibrium for comparison.

First time players who are matched against computer types br, t&e, or fic achieve median profits that are about equal to or slightly less than the Nash equilibrium profit. Drastically different, however, are profits of subjects who were matched against the computer types imi and re. Median profits against imi were less than half the profits against the first three computer types. Even the very best subjects do not reach the Nash
equilibrium profit, despite the bias in the noise of this computer type (see Footnote 7). Profits against computer type re are also substantially lower than against br and fic but they are higher than against imi. The range of profits is highest against this type of computer. Some subjects achieve very high profits that exceed the Stackelberg leader or collusive profit (of 1,458).

Median profits of repeaters are generally higher than those of first time players. While subjects improve somewhat against computer type imi, median profits are still by far the lowest of all computer types. Against br, fic, and re subjects achieve higher median profits than in the Nash equilibrium. Again, the very best subjects played against t&e and re.

It is also quite instructive to consider average profits over time. Figure 2 shows average profits of subjects and computers for all 40 periods. Subjects playing against type br almost immediately gain a substantive edge over the computer and keep their profits more or less constant somewhere between the Stackelberg leader profit and the Nash equilibrium profit. The final result against type fic is similar but convergence is much more gradual. This shows a considerable amount of foresight on the side of our subjects. When playing against fic (in contrast to br), subjects must be more patient and forward looking to “teach” the computer into a Stackelberg follower position. The

---

16 For first-time players, profits against re are lower than against br and fic according to two-sided, non-parametric Mann–Whitney U tests (see, e.g., Siegel and Castellan 1988) at $p < 0.001$. For repeaters these differences are not significant anymore. For both, first-timers and repeaters, profits against imi are lower than against any other computer type at $p < 0.001$. A robust rank order test yields the same significance levels.

17 Profits of those first-timers who we could identify as subsequent repeaters were actually lower in their first play than those of other first-timers although this difference is not significant. Thus, the increase in profits shown in Fig. 1 appears to be driven by experience rather than selection of subjects.
Fig. 2  Time series of profits for subjects and computers for different computer types

fictitious play computer is also the most successful among the computer types as it stabilizes at a profit of above 1,000. The time series against types t&e and re look similar, although against the latter subjects do not even manage to achieve the Nash equilibrium profit on average.\textsuperscript{18}

\textsuperscript{18} The dip of the computers’ profits in round 2 is due to the high relative weight of the (uniformly distributed) initial weights in early rounds, while the computer quantity in round 1 is not chosen by the learning theory, but set to 35, 40 or 45.
Computer type imi yields a totally different picture. In contrast to all others, it is the only computer type where subjects’ payoffs are lower than those of computers. Furthermore, subjects average profits are much lower than against other learning algorithms. Finally, the data show that payoffs against imi decrease over time, both for subjects and for computers. All three results are in line with the theoretical predictions of Proposition 1.

Table 3 considers the subjects with the highest profits overall. Among the top 100 subjects, there are 48 subjects who played against a computer of type re, 28 who played against type br, and 24 who played against t&e. The top ten players were almost exclusively playing against type re. This confirms the impression obtained from Fig. 1. The highest profits can be achieved against type re but a lot of luck is needed for this due to the stochastic nature of reinforcement learning.

5.2 Human tactics

In this section we shall answer the second question raised in the introduction, namely how can the opponent’s learning process be influenced by the player’s behavior. One particularly intriguing question concerns evidence for strategic teaching, i.e., strategic manipulation of the computer types by subjects. Subjects may forgo short-term gains in order to manipulate the programmed opponent and earn large gains in later periods.\(^\text{19}\)

Since initially, first-timers did not know the computer type, they may experiment with different quantities in order to explore the programmed opponent’s responses. Table 4 reports the fraction of subjects that experiment with quantities. We call a subject experimenting if the standard deviation of his quantities in the first 10 rounds is at least twice the standard deviation in the last 30 rounds. Overall at least one quarter of the first-time subjects experiment with quantities in this sense. The fraction exceeds 40% for fictitious play and trial & error learning. Note that these two are the “slowest moving” computer types. Table 4 also reports the fraction of repeaters who experiment. Interestingly, exploration declines when subjects play repeatedly except for reinforcement learning. So for all learning theories except reinforcement learning, exploration of first-timers may yield information that is used when the experiment is repeated. There may be two reasons for why it is different for reinforcement learning. First, note that reinforcement learning involves a probabilistic choice rule and

\(^{19}\) Collusion as an outcome is theoretically possible only against computer type t&e (Huck et al. 2004b). However, as data on individual plays reveal, there were no successful examples of collusion between subject and computer over a prolonged period.
Table 4  Classification of tactics

| Against computer type... | Tactic     | First-timer (%) | Repeater (%) |
|--------------------------|------------|-----------------|--------------|
| br                       | Experimentation | 25              | 18           |
|                          | Leadership  | 20              | 32           |
| fic                      | Experimentation | 41              | 29           |
|                          | Leadership  | 14              | 11           |
| re                       | Experimentation | 25              | 32           |
|                          | Leadership  | 16              | 16           |
| ini                      | Experimentation | 31              | 25           |
|                          | Leadership  | 16              | 15           |
| t&e                      | Experimentation | 40              | 18           |
|                          | Leadership  | 16              | 28           |

may appear quite erratic to subjects. Therefore it may take more effort to learn about reinforcement learning than about other computer types. Second, as we have seen in previous sections, with some luck subjects can earn large profits if reinforcement starts with low quantities. Subjects’ experimentation in the first 10 rounds may be aimed exactly at this.

Once subjects have explored and learned about the computer type, they may use this information to actively manipulate the computer type. Such manipulations may take on various forms. Probably the most straightforward form of manipulation is aimed at achieving Stackelberg leadership through aggressive play of large quantities. Table 4 also reports the fraction of subjects with such leadership behavior. We define a subject as displaying leadership behavior if he chooses a quantity of least 50 for at least 36 out of 40 rounds. About 16% of the first-timers display such leadership behavior. When playing against best response or trial & error learning, this behavior becomes even more pronounced among repeaters. The increase in leadership behavior is most remarkable when subjects play against br. Indeed, playing aggressively is a quite successful manipulation of br. Figure 3a shows quantities of the most successful subject playing against br and the corresponding computer quantities. This subject (ranked overall 45th) chose 55 in all 40 periods.20 The computer quickly adjusted to a neighborhood of the Stackelberg follower quantity with the remaining movement due to the noise in the computer’s decision rule.

While leadership may be a relatively simple form of strategic manipulation, individual data reveal manipulations that can be very sophisticated. We discovered quite interesting, though not very frequent, patterns that can be seen in Fig. 3b. The subject who played against best response chose—with only slight variations—the following cycle of 4 quantities: 108, 70, 54, 42, 108, 70, …Stunningly, this cycle produces an expected profit per round of 1,520, which exceeds the Stackelberg leader profit.21 By

---

20 Curiously, none of our subjects chose the exact Stackelberg leader quantity of 54.
21 The only reason the subject in Fig. 3a received an even higher payoff was luck due to favorable noise of the computer algorithm.
flooding the market with a quantity of 108, the subject made sure that the computer left the market in the next period. But instead of going for the monopoly profit, the subject accumulated intermediate profits over three periods. This, of course, raises the question, whether a cycle is optimal and how the optimal cycle looks like. It turns out, that in this game a cycle of length is four is optimal and, after rounding to integers, the optimal cycle is 108, 68, 54, 41, which produces an expected profit of 1,522.\textsuperscript{22} Thus, our subject was within 2 units of the solution for this non-trivial optimization problem.\textsuperscript{23}

\textsuperscript{22} See Schipper (2006) for a proof of this claim.

\textsuperscript{23} The subject played three times against br and left two comments. The first was “tried to trick him”, the second “tricked him”.

---

\textbf{Fig. 3} a Quantities of subject ranked number 45 and of the br-computer opponent (\textit{top panel}); b quantities of subject ranked number 49 and of the br-computer opponent (\textit{middle panel}); c quantities of top-ranked subject and his re-computer opponent
How did the very best subject play? Like all top players, he played against computer type re. Figure 3c reveals that the subject simply got lucky. The reinforcement algorithm locked in at very low quantities in the range between 10 and 20, and the subject roughly played a best response to that, which resulted in an average profit of 2,091.

One benchmark to compare the behavior of our subjects to is the maximal profit an omniscient, myopic player could achieve against the respective learning theory. To generate this benchmark, we ran simulations pitting our five computer types against a simulated player who can perfectly forecast the action his computer opponent is about to take (including the noise) and plays a best response to that, but disregards the influence of his action on the future behavior of his opponent. As Fig. 4 shows, our repeater subjects outperform that benchmark against br, re, and t&e. They do worse than the benchmark against fictitious play but considerably worse only against imitate the best. Given that the myopic best behavior requires a huge amount of knowledge about the opponent, which our subjects can not possibly possess since each learning
theory incorporates a random element, the only way for subjects to match or outperform the best myopic benchmark is by playing more sophisticated than myopic—by influencing future play of the learning theories.

6 The role of incentives: a robustness check

In the main (internet) experiment, the only incentive for subject was the highscore. It is a justified question whether these incentives are strong enough compared to the usual financial incentives given in laboratory experiments. To check for this, we conducted a control experiment as a regular laboratory experiment with the usual monetary incentives. In the lab experiment, 50 subjects played in the Bonn Laboratory for Experimental Economics. The instructions and the computer interface for both settings were the same up to the incentive structure. Subjects were required to repeat the experiment once with the same computer type as opponent, i.e., they played two times 40 rounds. Since there were fewer observations in the lab, we used only a starting value of 40 for the computer types. Incentives were provided by paying subjects immediately at the end of the experiment the sum of profits over all rounds according to an exchange rate of 9,000 Points to 1 euro. On average, subjects earned 10.17 euro for about half an hour in the lab.

We do find some significant differences between the two incentive structures. In the lab experiment, average quantities are significantly lower (MWU-test, $p < 0.01$) although average profits do not significantly differ. However, the crucial point is whether the differences across our treatments are robust to changing the incentive structure. The left panel of Fig. 5 shows average profits of all subjects given their respective computer opponent. The only significant difference between the lab and the internet is for computer opponent imi (MWU-test, $p < 0.05$). More importantly, all treatment effects are qualitatively the same, independently of the incentive structure. The right panel of Fig. 5 shows the average difference of the subject’s profit and the computer’s profit. The main result, that human subjects manage to exploit all of their computer opponents except imi, holds for both incentive structures. The average profit differential is significantly larger than 0 for computer opponents br, fic, re, and t&c and for both incentive structures, net and lab. The profit differential against imi is significantly smaller than 0 for both incentive structures at the 1% level. Furthermore, testing across incentive structures, profit differentials against br, re, and imi are not significantly different between net and lab. Profit differentials against fic

---

24 Recall that in the internet experiment, computer algorithms had an equal chance of starting with values 35, 40 or 45.

25 Both results hold for all subjects and for first-timers only. All tests in this Section can also be conducted as Kolmogorov–Smirnov tests without changing the results. For additional tables with data for each learning algorithm see the working paper version (Duersch et al. 2008a).

26 Likewise, when we recalculate Tables 3 and 4 by including data from the lab, only very minor changes occur.

27 According to $t$-tests at the 1% level with the exception of t&e in lab for which the differential is only marginally larger at the 10% level. Results for a Wilcoxon test are qualitatively the same.
and t&c are significantly larger in net but, qualitatively, the treatment effects seem to be robust.

7 Conclusion

In this experiment we let subjects play against computers which were programmed to follow one out of a set of popular learning theories. The aim was to find out whether subjects were able to exploit those learning algorithms. We find that there are remarkable differences in the exploitability of myopic learning algorithms. There are two insights from this observation: First, while the bulk of traditional learning theories that have been studied in the literature are myopic in nature, we need more advanced learning theories that incorporate at least a limited amount of foresight if those learning theories should also fit subjects engaged in strategic teaching. Many of our subjects were quite able to exploit the simple myopic learning algorithms. Strategic teaching is an important phenomenon that needs to be accounted for in the future development of theory.

Second, from an evolutionary perspective, one desideratum to impose on a realistic learning theory should be its non-exploitability. If a learning theory can be exploited by sophisticated subjects to their advantage, then such a learning theory should disappear in the long run. Interestingly, we find that among the learning algorithms studied in this paper, only imitate-the-best is robust to exploitation. This learning algorithm is known to lead to a non-Nash equilibrium in the game studied. This observation poses the following theoretical question left for further research: Does there exist a simple adaptive learning algorithm that is not exploitable and leads to Nash equilibrium?
Appendix A: Proof of Proposition 1

Proof Let the computer be player 2. We first claim that imi will imitate player 1’s quantity from the previous round if and only if it is closer to 54 than its own quantity and \(q_1 + q_2 \neq 108\). For \(q_1 + q_2 = 108\), both profits are zero and imi sticks to its own strategy. For \(q_1 + q_2 \neq 108\), we need to show that \(\pi(q_1, q_2) > \pi(q_2, q_1)\) if and only if \(|q_1 - 54| < |q_2 - 54|\). For \(q_1 + q_2 < 108\) or for \(q_1, q_2 > 54\), the claim follows because for \(p > c\) (\(p < c\)) the firm with higher (lower) quantity makes the larger profit (the smaller loss). It remains to consider the case \(q_i > 54 > q_{-i}\) but \(q_1 + q_2 > 108\). In this case

\[|q_i - 54| = q_i - 54 > 54 - q_{-i} = |q_{-i} - 54|.\]

Since \(p < c\), both firms make losses and since \(q_i > q_{-i}\), firm \(i\)'s losses are larger, which proves the claim. Note in particular, that as long as \(q_1 + q_2 < 108\), imi will never lower its quantity \(q_2\).

(a) Since quantities are strategic substitutes, the above claim implies that for starting values of the computer higher than the Cournot quantity, \(q_2^1 > 36\), player 1 is best off by not triggering imi to choose higher quantities than \(q_2^1\). Thus, imi will continue to play \(q_2^1\) and player 1’s optimal strategy is to play a myopic best reply \(q_1 \in \text{br}(q_2^1)\) every period. The resulting profits for the human player over 40 rounds are \(\max_{q_1} \pi(q_1, 40) = 46,240\) and \(\max_{q_1} \pi(q_1, 45) = 39,680\). The resulting profits for computer imi are \(\pi(40, q_1) = 54,440\) and \(\pi(45, q_1) = 57,600\).

If \(q_2^1 = 35\), player 1’s optimal strategy is to play \(q_1^t = q_2^1\), for all \(t = 1, \ldots, 39\) and \(q_1^{40} \in \text{br}(q_2^1)\). This results in a profit of 53,202. The imi computer’s profit is 53,165 in this case. Since the three starting values 35, 40, and 45 occur with the same frequency, the average profit for the player from this optimal strategy is 46,374, while his computer opponent imi earns 55,068.

(b) Given a starting value \(q_2\) of the computer, the player can obtain a positive profit differential in a given round \(t\) if and only if \(|q_1^t - 54| < |q_2^t - 54|\). Unless the player changes his strategy again, this profit differential is eroded in the next period due to the adjustment of imi to the quantity that yielded the higher profit. The player can close the gap to 54 in one large step or in several small steps. Due to the linear structure of the Cournot game, the number of steps does not matter. To see this note that the profit differential for one step is given by \((p - 1)(q_1 - q_2)\). For starting values of imi below 54, this expression is maximized by \(q_1 = 54\). Now suppose the player takes two arbitrary steps in rounds 1 and 2 to reach 54 (for all rounds thereafter the profit differential is zero). The profit differential is then given by

\[(p^1 - 1)(q_1^1 - q_2^1) + (p^2 - 1)(q_1^2 - q_2^2).\]

Since imi always imitates quantities that are closer to 54, we have that \(q_2^2 = q_1^1\). Thus \((5)\) is maximized by \(q_1^1 = 54\) and arbitrary \(q_1^2\) such that \(q_2^2 \leq q_1^2 \leq 54\). Consequently, the profit differential is the same with one or two steps. The same argument holds for any number of steps towards 54.
The maximal profit differentials that can be obtained for the three starting values 35, 40, and 45 are therefore 361, 196, and 81, respectively, which yield an average profit differential of 212.67.

Appendix B: Instructions

B.1 Introduction page

Welcome to our experiment!

Please take your time to read this short introduction. The experiment lasts for 40 rounds. At the end, there is a high score showing the rankings of all participants. You represent a firm which produces and sells a certain product. There is one other firm that produces and sells the same product. You must decide how much to produce in each round. The capacity of your factory allows you to produce between 0 and 110 units each round. Production costs are 1 per unit. The price you obtain for each sold unit may vary between 0 and 109 and is determined as follows. The higher the combined output of you and the other firm, the lower the price. To be precise, the price falls by 1 for each additional unit supplied. The profit you make per unit equals the price minus production cost of 1. Note that you make a loss if the price is 0. Your profit in a given round equals the profit per unit times your output, i.e., profit = (price − 1) * Your output. Please look for an example here. At the beginning of each round, all prior decisions and profits are shown. The other firm is always played by a computer program. The computer uses a fixed algorithm to calculate its output which may depend on a number of things but it cannot observe your output from the current round before making its decision. Your profits from all 40 rounds will be added up to calculate your high score. There is an overall high score and a separate one for each type of computer. Please do not use the browser buttons (back, forward) during the game, and do not click twice on the go button, it may take a short while.

Choose new quantity

Please choose an integer (whole number) between 0 and 110.

B.2 Example page

The Formula

The profit in each round is calculated according to the following formula:

Profit = (Price − 1) × Your output

The price, in turn, is calculated as follows.

Price = 109 − Combined output

That is, if either you or the computer raises the output by 1, the price falls by 1 for both of you. (but note that the price cannot become negative). And the combined output is simply:

Combined output = Your output + Computers output
Example:
Let's say your output is 20, and the computer's output is 40. Hence, combined output is 60 and the price would be 49 \((=109-60)\). Your profit would be \((49 - 1) \times 20 = 960\). The computer's profit would be \((49 - 1) \times 40 = 1,920\). Now assume you raise your output to 30, while the computer stays at 40. The new price would be 39 \((=109-40-30)\). Your profit would be \((39 - 1) \times 30 = 1,140\). The computer's profit would be \((39 - 1) \times 40 = 1,520\).

To continue, please close this window.

Appendix C: Screenshots
Welcome to our experiment!

Please choose a nickname for the highscore entry and please fill out correctly whether you played here before.

We will be online until mid of February.

Take a quick look at the highscore.

Contact us:
game@uni-bonn.de

---

**Round number 15 of 40**

Choose new quantity

Please choose an integer (whole number) between 0 and 110.

| Round | Your quantity | Your profit | Computer quantity | Computer profit |
|-------|---------------|-------------|-------------------|-----------------|
| 14    | 1             | 34          | 1360              | 1360            |
| 13    | 45            | 2205        | 14                | 686             |
| 12    | 75            | -75         | 39                | -39             |
| 11    | 27            | 1242        | 35                | 1610            |
| 10    | 38            | 1406        | 33                | 1221            |
| 9     | 48            | 1440        | 30                | 900             |
| 8     | 47            | 1551        | 20                | 924             |
| 7     | 46            | 1580        | 32                | 960             |
| 6     | 45            | 1350        | 33                | 990             |
| 5     | 45            | 1530        | 29                | 986             |
| 4     | 44            | 1540        | 32                | 1015            |
| 3     | 44            | 1400        | 32                | 1024            |
| 2     | 44            | 1452        | 31                | 1023            |
| 1     | 44            | 1276        | 35                | 1015            |

---

**Round number 1 of 40**

Round Your quantity Your profit Computer quantity Computer profit

| Round | Your quantity | Your profit | Computer quantity | Computer profit |
|-------|---------------|-------------|-------------------|-----------------|
| 1     | 44            | 1276        | 35                | 1015            |

---
Open Access This article is distributed under the terms of the Creative Commons Attribution Noncommercial License which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

References

Alós-Ferrer, C., Ania, A.B.: The evolutionary stability of perfectly competitive behavior. Econ Theory 26, 497–516 (2005)
Apesteguia, J., Huck, S., Oechssler, J.: Imitation: Theory and experimental evidence. J Econ Theory 136, 217–235 (2007)
Brown, G.W.: Iterative solutions of games by fictitious play. In: Koopmans, T.C. (ed.) Activity Analysis of Production and Allocation. Wiley, New York (1951)
Camerer, C., Ho, T.H.: Strategic learning and teaching in games. In: Hoch, S., Kunreuther, H. (eds.) Wharton on Decision Making. Wiley, New York (2001)
Camerer, C., Ho, T.H., Chong, J.K.: Sophisticated experience-weighted attraction Learning and strategic teaching in repeated games. J Econ Theory 104, 137–188 (2002)
Cason, T., Sharma, T.: Recommended play and correlated equilibria: An experimental study. Econ Theory 33, 11–27 (2007)
Coricelli, G.: Strategic interaction in iterated zero-sum games, Mimeo (2005)
Cournnot, A.: Researches into the mathematical principles of the theory of wealth, transl. by Bacon, N.T., MacMillan Company, New York, 1927 (1838)
Drehmann, M., Oechssler, J., Roider, A.: Herding and contrarian behavior in financial markets. Am Econ Rev 95(5), 1403–1426 (2005)
Duersch, P., Kolb, A., Oechssler, J., Schipper, B.: Rage against the machines—How Subjects Learn to Play Against Computers, AWI-Discussion Paper No. 423, Department of Economics, University of Heidelberg (2008a)
Duersch, P., Oechssler, J., Schipper, B.: Experimenting on the internet: Does it make a difference? University of Heidelberg, Mimeo (2008b)
Ellison, G.: Learning from personal experience: One rational guy and the justification of myopia. Game Econ Behav 19, 180–210 (1997)
Erev, I., Haruvy, E.: Learning and the Economics of Small Decisions. In: Kagel, J.H., Roth, A.E. (eds.) The Handbook of Experimental Economics, vol. 2 (forthcoming) (2008)
Fudenberg, D., Levine, D.K.: Reputations and equilibrium selection in games with a patient player. Econometrica 57, 759–778 (1989)
Fudenberg, D., Levine, D.K.: The Theory of Learning in Games. Cambridge: MIT Press (1998)
Huck, S., Normann, H.T., Oechssler, J.: Learning in Cournot oligopoly: An experiment. Econ J 109, C80–C95 (1999)
Huck, S., Normann, H.T., Oechssler, J.: Through trial & error to collusion. Int Econ Rev 45, 205–224 (2004a)
Huck, S., Normann, H.T., Oechssler, J.: Two are few and four are many: Number effects in experimental oligopoly. J Econ Behav Organ 53, 435–446 (2004b)
Ianni, A.: Reinforcement learning and the power law of practice: Some analytical results. Southampton: University of Southampton (2002)
Laslier, J.-F., Topol, R., Walliser, B.: A behavioral learning process in games. Game Econ Behav 37, 340–366 (2001)
Mosteriner, D., Shapley, L.: Potential games. Game Econ Behav 14, 124–143 (1996)
Offerman, T., Potters, J., Sonnemans, J.: Imitation and belief learning in an oligopoly experiment. Rev Econ Studies 69, 973–997 (2002)
Possajennikov, A.: Evolutionary foundation of aggregative-taking behavior. Econ Theory 21, 921–928 (2003)
Robinson, J.: An iterative method of solving games. Ann Math 54, 296–301 (1951)
Roth, A., Erev, I.: Learning in extensive form games: Experimental data and simple dynamic models in the intermediate term. Game Econ Behav 8, 164–212 (1995)
Sarin, R., Vahid, F.: Strategic similarity and coordination. Econ J 114, 506–527 (2004)
Schipper, B.C.: Submodularity and the evolution of Walrasian behavior. Int J Game Theory 32, 471–477 (2003)
Schipper, B.C.: Strategic control of myopic best reply in repeated games. Davis: University of California (2006)
Schipper, B.C.: Imitators and optimizers in Cournot oligopoly. Davis: University of California (2008)
Selten, R., Buchta, J.: Experimental sealed bid first price auctions with directly observed bid functions.
In: Budescu, D., Erev, I., Zwick, R. (eds.) Games and human behavior: Essays in honor of Amnon
Rapoport. Lawrence Erlbaum Associates, Mahwah (1998)
Shachat, J., Swarthout, J.T.: Learning about learning in games through experimental control of strategic
independence. Arizona: University of Arizona (2002)
Siegel, S., Castellan, N.J. Jr: Nonparametric Statistics for the Behavioral Sciences. Singapore: McGraw-
Hill (1988)
Tanaka, Y.: Long run equilibria in an asymmetric oligopoly. Econ Theory 14, 705–715 (1999)
Vega-Redondo, F.: The evolution of Walrasian behavior. Econometrica 65, 375–384 (1997)