Strategic Teaching and Learning in Games

Burkhard C. Schipper*

September 20, 2021

Abstract

It is known that there are uncoupled learning heuristics leading to Nash equilibrium in all finite games. Why should players use such learning heuristics? We show that there is no uncoupled learning heuristic leading to stage-game Nash equilibrium in all finite games that a player has an incentive to adopt, that would be evolutionary stable or that could "learn itself". Rather, a player has an incentive to strategically teach such a learning opponent in order to secure at least the Stackelberg leader payoff. The result remains intact when restricted to the classes of generic games, two-player games, potential games, games with strategic complements or 2×2 games, in which learning is known to be "nice". More generally, it also applies to uncoupled learning heuristics leading to correlated equilibria, rationalizable outcomes, iterated admissible outcomes, or minimal curb sets. A possibility result restricted to "strategically trivial" games fails if some generic games outside this class are considered as well.

Keywords: Interactive learning, uncoupled learning, equilibrium, repeated games, reputation, meta-learning.

JEL-Classifications: C72, C73.

*Department of Economics, University of California, Davis, Email: bcschipper@ucdavis.edu

Part of the research was undertaken when the author visited New York University and the University of Heidelberg. I thank the editor, three anonymous reviewers, Yakov Babicheko, Jerney Copic, Peter Duersch, Drew Fudenberg, Hans Haller, Martin Meier, Ichiro Obara, Joerg Oechssler, Joe Ostroy, Tomasz Sadzik, Klaus Schmidt, and Dale Stahl as well as seminar audiences at Queen's University, UC Davis, UCLA, USC, UT Austin, U Wisconsin-Madison, Virgina Tech, UN Reno, SAET Paris 2013, and GAMES Maastricht 2016 for helpful comments. Financial support through NSF CCF-1101226 is gratefully acknowledged.

1 Introduction

Individuals are not born with complete knowledge but with an ability to learn. In economic and social situations, learning is interactive as players learn about the behavior of other learners. There is now a large literature that studies how individuals learn equilibrium interactively (see for an overview Fudenberg and Levine, 1998a, or Young, 2004).

This paper is about the foundation of learning heuristics. Learning takes place in a repeated context. So it is natural to ask whether learning players not only reach equilibrium of the stage-game but also have a strategic incentive to adopt the learning heuristic in the long-run. Similarly, in an evolutionary context one may view players as programmed to learning heuristics. The question is then whether learning heuristics that converge to equilibrium of the stage-game can be evolutionary stable. Finally, one may allow players to learn about learning heuristics. The question becomes then whether there is a learning heuristic that could learn itself when applied to the game of choosing learning heuristics.

As in the most recent literature on learning, we focus on *uncoupled* learning heuristics that if followed by all players lead to a stage-game equilibrium in *all* games (e.g., Foster and Young, 2003, 2006, Hart and Mas-Colell, 2003, 2006, Germano and Lugosi, 2007, Kakade and Foster, 2008, Young, 2009). Our main observation is that when we want to find a learning heuristics that players have an incentive to adopt, can be evolutionary stable, or could learn itself, we need to look beyond uncoupled learning heuristics converging to equilibrium in all games.

The setting we study is as follows: Players face repeatedly a finite strategic game. Each player observes the past actions of all other players (i.e., perfect monitoring). Each player uses a learning heuristic, a strategy that assigns to each strategic game and each history in the repeated version of that game a mixed action. The literature on learning in games has focused on learning heuristics that satisfy at least the following two requirements: First, they should lead to Nash equilibrium in every finite stage game. That is, when all players adopt such a learning heuristic, their behavior should eventually converge to Nash equilibrium of the stage game. Second, learning heuristics should be uncoupled, i.e., they should not directly take opponents' payoffs as input.

The uncoupledness assumption requires some discussion. In game theory, structural assumptions like uncertainty about payoffs are typically modelled within the game while behavioral assumptions are modelled with conditions on strategies. In the literature of learning, this conceptual divide is blurred as uncoupledness has been interpreted as complete ignorance about opponents' payoffs even though we have a well-established apparatus in game theory for modelling incomplete information about opponents' payoffs. In this paper, we treat uncoupledness as how it is formally modelled, namely as a behavioral assumption on strategies. As game theory attempts to explain behavior given the structural assumptions on the game, we find it natural to ask whether there exist strategic, evolutionary, or learning explanations for the use

uncoupled learning heuristics given the class of repeated finite games with complete information.

To answer the question, we impose a third requirement on learning heuristics. Consider the normal-form game associated with the repeated game in which actions are learning heuristics and payoffs are the long-run payoffs from profiles of learning heuristics. Since there are many games, we consider the "average" normal-form game defined by the expected long-run payoffs w.r.t. the Lebesgue measure on the space of games. We call this the learning game. Our requirement now is that a learning heuristic should also be a Nash equilibrium action of the learning game. The reason is that when a pair of learning heuristic is Nash equilibrium of the learning game, then each player does not want to deviate unilaterally from the profile learning heuristics. Nash equilibrium of the learning game is conceivably also a necessary condition for evolutionary stability of learning heuristics. Finally, being Nash equilibrium of the learning game is also a necessary condition for an equilibrium learning heuristic to select itself when applied to the learning game.

Unfortunately, we show by a simple counterexample that there is no learning heuristic that is uncoupled, converges to Nash equilibrium in every finite game, and is Nash equilibrium of the learning game. This remains true even if instead of convergence to stage-game Nash equilibrium in *all* finite games, we just require convergence to Nash equilibrium in the class of 2×2 games, two-player games, potential games, games with strategic complements etc. in which learning is known to be "nice". The recent literature on learning showed that convergence to Nash equilibrium in all finite games is considerably more demanding than convergence to just correlated equilibrium. Yet, our counterexample also demonstrates that there is no learning heuristic that is uncoupled, converges to correlated equilibrium in every finite game, and is Nash equilibrium of the learning game. This remains true even if we consider convergence to iterated admissible action profiles, rationalizable action profiles or minimal curb-sets.

We show that for any uncoupled learning heuristic that converges to Nash equilibrium in all finite stage-games, there is a strategic teaching heuristic that can manipulate the learning opponent in such a way as to obtain the Stackelberg leader payoff "averaged" over all games. This is reminiscent of the reputation results in repeated games (e.g., Fudenberg and Levine, 1989). Strategic teaching has been observed in the experimental literature on learning (Duersch et al, 2010, Terracol and Vaksmann, 2009, Chong et al., 2006, Camerer et al., 2002, Hyndman et al., 2012). The fact that long-run payoffs strictly larger than stage-game Nash equilibrium payoffs can be earned against uncoupled equilibrium learners suggests that there are strict positive incentives for learning strategies that also feature espionage of opponent's payoffs and thus go beyond uncoupledness.

In the setting briefly described so far we allowed deviations with coupled learning heuristics. This makes sense because want to check whether we can derive uncoupledness of learning heuristics as equilibrium property of the learning game rather than imposing it as an assumption. Nevertheless, it begs the question whether there would be learning heuristics converging

to stage-game Nash equilibrium in all finite games that are also Nash equilibrium in the learning game *when players are restricted to uncoupled strategies only.* We extend our counterexample to show that there is no stationary uncoupled learning heuristic with finite recall converging to stage-game Nash equilibrium in every finite game and being a Nash equilibrium of the learning game. The reason is that if the opponent uses a stationary uncoupled learning heuristic with finite recall converging to stage-game Nash equilibrium, we can find an uncoupled learning heuristic that can first learn about the payoffs from the opponent's behavior and then use this information to strategically teach the opponent to her advantage. For this to work, we allow strategic teaching heuristics with arbitrary long recall. Thus, we essentially replace coupled deviations by uncoupled deviations with large recall. For every uncoupled equilibrium learning heuristic with finite recall there is strict incentive for acquiring larger recall and use this to implicitly learn about opponent's payoffs.

Our counterexample begs the question whether there is a "maximal" class of games for which there are uncoupled learning heuristics that lead to Nash equilibrium in all games of this class, that each player has an incentive to adopt if opponents adopt their part, and for which this possibility fails the moment some other games outside the class are considered as well. We show that when we restrict to the class of games that can be solved by one round of elimination of weakly dominated actions or to the class of common interest games, then we obtain possibility results. Note though that these two classes of games are in some sense "strategically trivial" as players can deduce some of their own stage-game Nash equilibrium actions just from their own payoffs.

The next section outlines the model. The counterexample is presented in Section 3. The generality of the example is explored in Section 4. In Section 5 we establish lower bounds on payoffs achievable with strategic teaching. Some "possibility" results are presented in Section 6. In Section 7 we extend our observations to uncoupled strategic teaching heuristics. Section 8 explores discounting payoffs, games with mixed equilibrium only, and $(1-\varepsilon)$-convergence to stage-game ε-equilibrium. We conclude with a discussion in Section 9. Proofs are elementary and collected in the appendix.

2 Basic Model

For our purpose it is enough to consider two-player games with players Rowena R and Colin C. For simplicity, each player has the same nonempty finite set of actions A. As customary, $i \in \{R, C\}$ refers to one player and $-i \in \{R, C\}$ refers to i's opponent. The payoff function of player i is denoted by $u_i : A \times A \longrightarrow [0, 1]$. Later in the text, we slightly abuse notation and let u_i, $i \in \{R, C\}$, also denote the multilinear extended (expected) utility function defined on the space of mixed action profiles $\Delta(A) \times \Delta(A)$. We normalize payoffs to be in the unit interval for integrability reasons. Consider now the class of *all* two-player games in normal-form with

the action set A. The space of all such games is generated by varying players' payoff functions. Since A is finite, it is identified by the $2|A^2|$ - dimensional Euclidean vector space $[0,1]^{2|A^2|}$. Thus, we may simply call a profile of payoff functions $\mathbf{u} = (u_R, u_C) \in [0,1]^{2|A^2|}$ a game. Let λ be the Lebesgue measure on the space of games $[0,1]^{2|A^2|}$.[1] We let $\mathbf{U} \subseteq [0,1]^{2|A^2|}$ denote a measurable subset of games.

For each game $\mathbf{u} \in [0,1]^{2|A^2|}$, the dynamic setup consists of repeated play of the stage-game $\mathbf{u}$ at discrete time periods $t = 0, 1, 2, ...$ Let $a_i^t \in A$ denote the action of player i at time t, and $\mathbf{a}^t = (a_R^t, a_C^t) \in A \times A$ be the combination of actions at t. At the end of period t each player i observes the combination of realized actions $\mathbf{a}^t$. That is, we assume perfect monitoring of play.

A learning heuristic σ_i of player i assigns to each game $\mathbf{u} \in [0,1]^{2|A^2|}$ a sequence of functions $(\sigma_i^0(\mathbf{u}), \sigma_i^1(\mathbf{u}), ..., \sigma_i^t(\mathbf{u}), ...)$ where for each $t > 0$ the function $\sigma_i^t(\mathbf{u})$ assigns to each history $h^{t-1} := (\mathbf{a}^0, \mathbf{a}^1, ..., \mathbf{a}^{t-1})$ in the t-th repetition of the stage game $\mathbf{u}$ a mixed action[2] in $\Delta(A)$ to be played at stage t. This rather general definition is in line with the recent learning literature (e.g., Hart and Mas-Colell, 2006). With this formulation we let $\sigma_i^0(\mathbf{u})$ be simply player i's distribution of initial actions in game $\mathbf{u}$ when i follows learning heuristic σ_i. We assume that each σ_i^t, $t = 0, 1, ...$, is Lebesgue measurable with respect to the space of games. We denote by $\boldsymbol{\sigma} = (\sigma_R, \sigma_C)$ a profile of learning heuristics. The set of learning heuristics is denoted by Σ.

As discussed in the introduction, the literature on learning focused on uncoupled learning heuristics. These learning heuristics may take opponents' actions and the player's payoff function as an input but not opponents' payoff functions.

Definition 1 (Uncoupled) *A learning heuristic σ_i is uncoupled on a class of games $\mathbf{U} \subseteq [0,1]^{2|A^2|}$ if for every $(u_i, u_{-i}) \in \mathbf{U}$, $\sigma_i(u_i, u_{-i}) = \sigma_i(u_i, \hat{u}_{-i})$ holds for all $(u_i, \hat{u}_{-i}) \in \mathbf{U}$.*

Note that we explicitly define uncoupledness relative to a class of games. When we consider a class of games that is a strict subset of the set of all finite two-player games, then this weakens the uncoupledness assumption as we implicitly allow players to make use of the information that the opponent's payoffs are within this class of games.

The following notion of convergence to stage-game equilibrium can be viewed both as strong and weak. It is strong because we require almost sure convergence.[3] We believe that it captures

[1] The Lebesgue measure gives us a notion of "size" of various classes of games. Our results generalize to smooth measures.

[2] Allowing for mixed actions and thus stochastic learning heuristics is crucial (even for learning pure Nash equilibrium). Hart and Mas-Colell (2003) show that there exist no *deterministic* uncoupled learning heuristics converging to Nash equilibrium in all games while Hart and Mas-Colell (2006) show that there are *stochastic* uncoupled learning heuristics that converge to Nash equilibrium in all games. Randomization of actions allows for exhaustive search.

[3] To keep the exposition simple and conceptually straightforward, we use almost sure convergence. In Section 8.2 we generalize our result to a weaker notion of convergence.

best what we intuitively mean with "convergence" to equilibrium. Yet, it is also a weak notion of convergence in the sense that we require convergence to pure equilibrium only and are silent on convergence in games with mixed equilibrium only (see Section 8.3). As it will become clear later, the focus on pure equilibrium will be enough for our purpose.

Definition 2 (Convergence to Nash Equilibrium) *A profile of learning heuristics $\boldsymbol{\sigma}$ converges to Nash equilibrium in every stage-game in $\mathbf{U} \subseteq [0,1]^{2|A^2|}$ if for every game $\mathbf{u} \in \mathbf{U}$ that possesses a pure Nash equilibrium, almost every play path consists of a pure stage-game Nash equilibrium being played from some point on. We say that learning heuristic σ_i converges to Nash equilibrium in every stage-game if it is a component of a profile of learning heuristics that converges to Nash equilibrium in every stage-game.*

An example of a learning heuristics that satisfies both uncoupledness and convergence to stage-game Nash equilibrium is the one used by Hart and Mas-Colell (2006, proof of Theorem 3): If each player plays the same action in the past two periods and player i's action is a best response to player $-i$'s action, then player i plays the same action again. Otherwise, each randomizes uniformly over all actions. Clearly, if players play equilibrium for two periods, they are stuck there. Otherwise, they may play equilibrium by chance in the next period and by chance in the next-next period in which case they get stuck.

While the simplistic "learning" heuristic just described is uncoupled and Nash convergent in any finite game with pure equilibrium, it is not clear that players have a long-run incentive to follow it. In order to check whether there could be any learning heuristic that is uncoupled, Nash convergent, and that players have a long-run incentive to adopt, we describe next a game in which the "actions" are the learning heuristics. We denote by $\mathbf{a}^t(\boldsymbol{\sigma}(\mathbf{u}))$ a profile of actions realized in period t under a history generated by the profile of learning heuristics $\boldsymbol{\sigma}$ in the repeated stage-game $\mathbf{u}$. For each profile of learning heuristics $\boldsymbol{\sigma}$ and each game $\mathbf{u} \in [0,1]^{2|A^2|}$, we denote player i's limit of expected means payoff by

$$v_i(\boldsymbol{\sigma}(\mathbf{u})) := \lim_{T\to\infty} \inf \mathbb{E}_{\boldsymbol{\sigma}(\mathbf{u})} \left[\frac{1}{T} \sum_{t=1}^{T} u_i(\mathbf{a}^t(\boldsymbol{\sigma}(\mathbf{u}))) \right], \tag{1}$$

where the expectations are formed over mean payoffs[4] resulting from histories given positive probability by the profile of learning heuristics $\boldsymbol{\sigma}$ in the repeated stage-game $\mathbf{u}$. This is the long run expected payoff to player i in game $\mathbf{u}$ emerging from the profile of learning heuristics $\boldsymbol{\sigma}$. Note that v_i is a measurable random variable on $[0,1]^{2|A^2|}$. To see this note that by assumption σ_i^t is measurable for every $t = 0, 1, ...$ Further, for every T, $\mathbb{E}_{\boldsymbol{\sigma}(\mathbf{u})} \left[\frac{1}{T} \sum_{t=1}^{T} u_i(\mathbf{a}^t(\boldsymbol{\sigma}(\mathbf{u}))) \right]$ is linear in probabilities of the per-period behavior strategies on a finite dimensional real-valued

[4]To keep the exposition simple and conceptually straightforward, we use the limit of expected means payoff over alternative ways to evaluate streams of payoffs. In Section 8.1 we show that our results extends to discounting for sufficiently patient players.

domain. Hence it is continuous in those probabilities and thus measurable on $[0,1]^{2|A^2|}$. Since the lim inf of a sequence of measurable real-valued functions is measurable, we have that v_i is a measurable random variable on $[0,1]^{2|A^2|}$.

As discussed in the introduction, we are interested in the long run expected payoff "averaged" over all games in a Lebesgue measurable subset of games $\mathbf{U} \subseteq [0,1]^{2|A^2|}$ defined by

$$V_i(\boldsymbol{\sigma}, \mathbf{U}) := \int_{\mathbf{U}} v_i(\boldsymbol{\sigma}(\mathbf{u}))d\lambda. \tag{2}$$

This defines a game in normal-form $\langle \{R, C\}, \Sigma, (V_i(\cdot, \mathbf{U}))_{i=R,C} \rangle$ in which each player i "chooses" a learning heuristic in Σ and her payoff from a profile of learning heuristics $\boldsymbol{\sigma} \in \Sigma \times \Sigma$ over all games in $\mathbf{U}$ is given by $V_i(\boldsymbol{\sigma}, \mathbf{U})$. We call this game the *learning game* based on $\mathbf{U}$.

We are interested in learning heuristics that are a Nash equilibrium of a learning game. Existence of Nash equilibrium of the learning game is guaranteed. There is a learning heuristic that for each stage-game $\mathbf{u} \in \mathbf{U}$ prescribes a Nash equilibrium of this stage-game. Of course, such a strategy would not necessarily be uncoupled except when considering just some special classes of games.

Definition 3 (Nash Equilibrium of the Learning Game) *A profile of learning heuristics $\boldsymbol{\sigma} = (\sigma_R, \sigma_C) \in \Sigma \times \Sigma$ is a Nash equilibrium of the learning game $\langle \{R, C\}, \Sigma, (V_i(\cdot, \mathbf{U}))_{i=R,C} \rangle$ if for $i \in \{R, C\}$,*

$$V_i(\sigma_i, \sigma_{-i}, \mathbf{U}) \geq V_i(\hat{\sigma}_i, \sigma_{-i}, \mathbf{U}) \text{ for all } \hat{\sigma}_i \in \Sigma.$$

As we explained in the introduction, we view this requirement as a weak necessary condition for a strategic, evolutionary, or learning foundation of learning heuristics.[5] Note that currently we allow for deviations with any learning heuristics in Σ. The extension to uncoupled deviations is deferred to Section 7.

3 A Simple Counterexample

Consider the following 2×2 game $\mathbf{u}^1$ defined by

		Colin	
		a	b
Rowena	b	$16, 12$	$13, 13$
	c	$17, 7$	$14, 6$

[5]Equilibrium among learning heuristics/rules is not an entirely new idea. Germano (2007) considers Nash equilibrium of boundedly rational rules for playing games. It is also similar to Nash equilibrium of learning heuristics game-by-game in Branfman and Tennenholtz (2004), Ashlagi et al. (2006), and Monderer and Tennenholtz (2007).

In this game, action c of Rowena strictly dominates her action b. The unique Nash equilibrium of the game is (c, a), which is in pure actions. If players follow uncoupled learning heuristics converging to Nash equilibrium in all finite games, then they must reach (c, a) in game $\mathbf{u}^1$.

Next consider the 2×2 game $\mathbf{u}^2$ defined by

		Colin	
		a	b
Rowena	b	$16, 12$	$13, 13$
	c	$15, 7$	$9, 6$

This game is identical to game $\mathbf{u}^1$ except for the payoffs of Rowena from action c. Now b strictly dominates action c. The unique Nash equilibrium is (b, b), which again is in pure actions. Any profile of learning heuristics that is uncoupled and converges to Nash equilibrium in all games must converge to (b, b).

Would Rowena have an incentive to stick to such a learning heuristic? Note that if Rowena behaves in game $\mathbf{u}^2$ exactly as in game $\mathbf{u}^1$, then since Colin follows an uncoupled learning heuristic the play must converge to (c, a) in game $\mathbf{u}^2$. This would yield Rowena a payoff of 15, which is strictly larger than her payoff of 13 in the unique Nash equilibrium of $\mathbf{u}^2$. Thus, Rowena has a strict incentive to deviate from an uncoupled learning heuristic converging to Nash equilibrium in all games to a "strategic teaching" heuristic as just described.

Since both games are generic, we can consider open neighborhoods $\mathbf{U}^1$ and $\mathbf{U}^2$ of $\mathbf{u}^1$ and $\mathbf{u}^2$, respectively, such that both $\mathbf{U}^1$ and $\mathbf{U}^2$ belong to the class of games $\mathbf{U}$ for which there exists a pure Nash equilibrium. Let (σ_R, σ_C) be a profile of uncoupled learning heuristics that converges to Nash equilibrium in games in $\mathbf{U}$. Moreover, let σ'_R be a learning heuristic that plays both in games in $\mathbf{U}^1$ and $\mathbf{U}^2$ identical to σ_R in $\mathbf{U}^1$ and identical to σ_R in all other games in $\mathbf{U}$. Since the space of games $\mathbf{U}$ contains the non-empty open sets $\mathbf{U}^1$ and $\mathbf{U}^2$ and the Lebesgue measure is strictly positive on nonempty open subsets of the space of games[6], the learning heuristic σ'_R is strictly better against σ_C in the learning game than σ_R against σ_C. We conclude: *There is no uncoupled learning heuristic that both converges to Nash equilibrium in all finite games and is a Nash equilibrium learning heuristic of the learning game.*

4 How General is the Counterexample?

Our arguments did not make reference to any other properties of learning rules that have been discussed in the literature such as m-periods of recall, m-memory, stationarity (Hart and Mas-Colell, 2006) or radical uncoupledness (Foster and Young, 2006, Germano and Lugosi, 2007, Young, 2009). Some of those properties are sometimes imposed on top of uncoupledness to

[6]Obviously, to fit the examples within our model outlined in the previous section, we would need to normalize payoffs of games in $\mathbf{U}^1$ and $\mathbf{U}^2$ with affine transformations to be within $[0, 1]$.

obtain positive or negative results on learning Nash equilibrium. Our observation holds for any such uncoupled learning heuristics converging to Nash equilibrium in all games.

There are simple learning heuristics approaching correlated equilibrium but not necessarily Nash equilibrium (Foster and Vohra, 1997, Fudenberg and Levine, 1999, and Hart and Mas-Colell, 2000, 2001, 2003). Note that in both games, $\mathbf{u}^1$ and $\mathbf{u}^2$, the unique correlated equilibrium is the unique Nash equilibrium. Similarly, in both games the unique Nash equilibrium coincides with the sets of iterated admissible strategy profiles, rationalizable strategy profiles (Spohn, 1982, Bernheim, 1984, Pearce, 1984), and minimal curb sets (Basu and Weibull, 1991).

Maybe asking for convergence in *all* games in $0, 1^{2|A^2|}$ is requiring too much. What if we restrict to generic and economically interesting class of games for which learning is known to be well-behaved? Note that our arguments make use of two-player games only. In fact, we use 2×2 games only. Moreover, both games $\mathbf{u}^1$ and $\mathbf{u}^2$ are ordinal potential games (in the sense of Monderer and Shapley, 1996), with ordinal potentials given, respectively, by

$$P^1 = \begin{array}{c} \\ b \\ c \end{array} \begin{array}{c} \begin{array}{cc} a & b \end{array} \\ \begin{pmatrix} 5 & 8 \\ 10 & 9 \end{pmatrix} \end{array} \qquad P^2 = \begin{array}{c} \\ b \\ c \end{array} \begin{array}{c} \begin{array}{cc} a & b \end{array} \\ \begin{pmatrix} 2 & 3 \\ 1 & -1 \end{pmatrix} \end{array}$$

Note further that both games satisfy strategic complements. That is, in both games and for each player there is an order on the set of actions such that both stage-games satisfy increasing differences, $u_i(x'', y') - u_i(x', y') \leq u_i(x'', y') - u_i(x', y'')$ for any actions $x'' > x'$, $y'' > y'$ and i, a property that facilitates convergence of many learning heuristics (e.g., Milgrom and Roberts, 1990). E.g., let $b < c$ for player 1 and $b < a$ for player 2. Both games we use have a unique Nash equilibrium which is in pure actions. Thus, the observation is not due to the fact of converging to a "wrong" equilibrium or miscoordination. Often stronger positive results on learning have been obtained for generic games only (see for instance Germano and Lugosi, 2007, Young, 2009).[7] Yet, all games we use are generic and "non-pathological". Thus, one cannot hope to obtain positive results for aforementioned subclasses of games. This seems particularly frustrating for the classes of 2×2 games, ordinal potential games, and games with strategic complements for which prominent learning heuristics are known to converge to equilibrium, see Fudenberg and Levine (1998a).

From the discussion, we conclude:

Proposition 1 *Let* $\mathbf{U}$ *be the class of all finite games. There is no learning heuristic that is uncoupled, converges to Nash equilibrium in all games in* $\mathbf{U}$ *and is a Nash equilibrium learning*

[7] The mathematical notion of genericity does not always coincide with economic relevance in game theory. The games used in our counterexample are relevant to economics. In fact, game $\mathbf{u}^2$ can be viewed as a textbook-style Cournot duopoly in which player i's profit function is $\max\{10.9 - q_i - q_{-i}, 0\} \cdot q_i - 0.1 \cdot q_i$, and for which we erase all quantities q_i and q_{-i} except for the Cournot Nash equilibrium quantity as well as the Stackelberg leader and follower quantities of each player and suitably rounded payoffs. That's why we use actions $\{b, c\}$ for Rowena and $\{a, c\}$ for Colin (rather than $\{U, D\}$ and $\{L, R\}$, respectively).

heuristic of the learning game based on $\mathbf{U}$. *This holds even if we restrict* $\mathbf{U}$ *to be the class of all generic finite games, all two-player games, all* 2×2 *games, all games with pure Nash equilibrium, all games with a unique Nash equilibrium, all games with strategic complements, all ordinal potential games, or any union thereof, or if we weaken convergence to correlated equilibrium, the set of iterated admissible strategy profiles, rationalizable strategy profiles, or minimal curb sets.*

Are all three properties necessary for the impossibility result? First, consider uncoupledness. For instance, the Lemke-Howson algorithm finds a Nash equilibrium in every finite two-player game of complete information (Lemke and Howson, 1964). It is a coupled algorithm that just takes payoff information as input and ignores behavior of the opponent. (Consider it as a strategy that at each period outputs the action of the profile of actions computed at that period and repeats the last action forever once the algorithm stops.) Thus, it cannot be strategically taught. Since the repeated stage-game equilibrium is an equilibrium of the repeated game, we also have that the Lemke-Howson algorithm is a Nash equilibrium of the learning game. Thus, uncoupledness is necessary for Proposition 1.

Proposition 1 applies to convergence of Nash equilibrium, correlated equilibrium, admissibility, rationalizability, and minimal CURB sets of the stage-game. All these solution concepts have in common that behavior convergences to a best response in the stage-game. Is this property necessary? For instance, rational learning à la Kalai and Lehrer (1993) with patient players converges (under suitable conditions on prior beliefs) to Nash equilibrium of the *repeated game* and hence to equilibrium of the learning game. If we consider rational learning as an uncoupled heuristics, then our counterexample shows that patient rational learning cannot converge to a best response of the stage-game. This would demonstrate that the requirement of convergence to stage-game best response is necessary for Proposition 1. However, the argument is not entirely satisfactory though since convergence to repeated games equilibrium requires some version of the "grain-of-truth" assumption, roughly requiring that each player's prior belief does not rule out the correct strategy of the opponent. Since the correct strategy is also rational for the opponent, indirectly each player takes some payoff information of the opponent into account. Thus, rational learning is not entirely uncoupled. So the question of whether convergence to stage-game best response is necessary for Proposition 1 is not entirely settled.

Finally, it is well-known that there are uncoupled learning heuristics converging to stage-game Nash equilibrium in all finite games (Foster and Young, 2003, 2006, Hart and Mas-Colell, 2003, 2006, Germano and Lugosi, 2007, Kakade and Foster, 2008, Young, 2009). Thus, the additional requirement that the learning heuristic is a Nash equilibrium of the learning game is necessary for Proposition 1.

5 Incentives for Strategic Teaching

In this section, we are interested in establishing a lower bound on the "average" long run payoffs that can be achieved against an opponent who adopted an uncoupled learning heuristic that converges to Nash equilibrium in all games. This allows us to draw a parallel to results in the literature on reputations in repeated games.

Define player i's pure best response correspondence for the stage-game $\mathbf{u} \in \mathbf{U}$ by

$$B_i(\mathbf{u})(a_{-i}) := \left\{ a_i \in A_i : u_i(a_i, a_{-i}) \geq u_i(a_i', a_{-i}) \text{ for all } a_i' \in A_i \right\}.$$

Further, define player i's worst Stackelberg leader payoff in game $\mathbf{u} \in \mathbf{U}$ by

$$\ell_i(\mathbf{u}) := \max_{a_i \in A_i} \min_{a_{-i} \in B_{-i}(\mathbf{u})(a_i)} u_i(a_i, a_{-i}).$$

In this definition, the Stackelberg leader is pessimistic since in case of multiple best responses of the follower, he assumes that the follower chooses the best response that is worst to him. This seems appropriate since our aim is to establish a lower bound. Yet, best responses are unique in generic games. Thus, for "average" long run payoffs (i.e., "averaged" over all games with respect to the Lebesgue measure on the space of games), the "pessimistic" selection from the follower's best response correspondence does not matter.

Let

$$L_i(\mathbf{U}) := \int_{\mathbf{U}} \ell_i(\mathbf{u}) d\lambda$$

be the "average" of Stackelberg leader payoffs over games in the Lebesgue measurable class $\mathbf{U}$ with respect to the Lebesgue measure λ.

A player who faces an opponent following an uncoupled learning heuristic that converges to Nash equilibrium in all finite two-player games that possess a pure Nash equilibrium can guarantee herself almost surely at least the Stackelberg leader payoff averaged over all games with a suitable strategic teaching strategy. Moreover, this payoff is strictly larger than adopting the uncoupled learning heuristic as well and converging to a Nash equilibrium in all games. In this sense, there is a strict positive incentive for strategic teaching to prevent learning from converging to Nash equilibrium in all games. This is stated more formally in the following proposition:[8]

Proposition 2 *Let $\mathbf{U}$ be the class of finite two-player games that possess a pure Nash equilibrium. For any profile of uncoupled learning heuristics (σ_R, σ_C) that converges to Nash equilibrium in all games in $\mathbf{U}$, there exists a learning heuristic $\tilde{\sigma}$ (i.e., a "strategic teaching heuristic")*

[8]Here and in the following sections Propositions 2 to 7 continue to hold when instead of requiring convergence to Nash equilibrium, we just require convergence to correlated equilibrium, iterated admissible profiles, rationalizable profiles, or minimal curb-sets.

such that when Rowena follows $\tilde{\sigma}$ and Colin follows σ_C, we have

$$V_R((\tilde{\sigma}, \sigma_C), \mathbf{U}) \geq L_R(\mathbf{U}) > V_R((\sigma_R, \sigma_C), \mathbf{U}).$$

The proof is contained in Appendix. We show that a player who is facing an opponent with an uncoupled learning heuristic converging to Nash equilibrium can guarantee herself with a "strategic teaching heuristic" at least the Stackelberg leader payoff in every generic two-player game that possesses a pure Nash equilibrium. This is because for every generic two-player game and Stackelberg outcome, there is a generic game in which this Stackelberg outcome is the unique Nash equilibrium. This game differs from the first one in the player's payoffs only but not in the opponent's payoffs. The player can then "pretend" to be in this game whenever the original game is played and strategically teach the opponent to jointly reach almost surely the outcome in the second game because the opponent follows an uncoupled learning heuristic. In this sense, the proof of Proposition 2 generalizes our counterexample as it shows that there is an opportunity for strategic teaching in *every* generic two-player game that possesses a pure Nash equilibrium. It does not follow yet that the strategic teacher earns a strictly higher payoff than in equilibrium of every game. The fact that the "average" over Stackelberg leader payoffs is strictly larger than the "average" limit-of-means payoffs of $\boldsymbol{\sigma}$ is perhaps not obvious. It is known that there are finite two-player games with a pure Nash equilibrium in which the Stackelberg leader payoff as defined here can be strictly below a Nash equilibrium payoff (e.g., Başar and Olsder, 1999, p. 132-133). But such games must be non-generic and do not have positive measure when "averaging" expected limit-of-means payoffs over all games in $\mathbf{U}$ with respect to λ. In generic games with pure equilibrium, the worst Stackelberg leader payoff is weakly above equilibrium payoff. Since our counterexample shows that there are also open subsets of games where the worst Stackelberg leader payoff is strictly above equilibrium payoff, the strict inequality in the proposition follows.

Proposition 2 is reminiscent of the reputation results in repeated games. For instance, Fudenberg and Levine (1989) consider repeated games with a long-run player who faces a sequence of short-run players. Short-run players are uncertain about the payoff-type of the long-run player and thus the game they are playing. In particular, there is strict positive prior probability that the long-run player is of the payoff-type for which the Stackelberg leader strategy of the "true" game is strictly dominant. They show that the long-run player's payoff converges to the Stackelberg leader payoff as she becomes sufficiently patient. In our setting, there is no Bayesian game. Yet, the role of uncertainty is taken by uncoupledness of the learning heuristic as it means that the learning heuristic cannot distinguish between "payoff-types" of the other player. Implicitly there are many "payoff-types" of the other player in our setting since we require an uncoupled learning heuristic to face the other player in *all* games with a pure equilibrium because it is supposed to converge to Nash equilibrium in all those games. The long-run player corresponds now to the strategic teacher who cares about her long-run payoff

that she can achieve against an opponent adopting an uncoupled learning heuristic leading to Nash equilibrium in all games. Moreover, the opponent adopting an uncoupled learning heuristic leading to Nash equilibrium in all games is short-term since he implicitly cares about converging to *stage-game* Nash equilibrium in all games. To some extent, Fudenberg and Levine (1998a, Chapter 8.11.1) anticipated our result in the last chapter of their book on "The Theory of Learning in Games" when they suggested that reputation results from the repeated games literature should carry over to the learning literature. Yet, they had a very different setting in mind in which they explicitly added uncertainty. Moreover, they did not show that any uncoupled learning heuristic that leads to Nash equilibrium provides an opportunity for strategic teaching and developing reputations.

6 "Possibility" Results

Uncoupled learning heuristics depend only on the player's own payoff function and the behavior of the opponent but not directly on the opponent's payoffs. With such learning heuristics "play has a decentralized character, and no player can, alone, recognize a Nash equilibrium" (Hart and Mas-Colell, 2006, p. 287). Thus, it is reasonable to expect that we could obtain a positive result when we restrict to the class of games in which both player's being rational and knowledge of their own payoffs implies equilibrium. In Nash equilibrium of such games, each player plays as if he solves individual decision problems. Nothing about the opponent's payoff function needs to be learned by the players in order to find the solution to the game. Hence, we may label such games "strategically trivial". We will show that indeed for such classes of games possibility results can be obtained but that these possibility results fail when any measurable subset of games with positive Lebesque measure outside these classes are considered as well.

We use as a rationality criterion the notion of admissibility, i.e., the avoidance of weakly dominated actions.

Definition 4 (Weak Dominance) *An action $a_i \in A$ is weakly dominated if there is a mixed action $\alpha_i \in \Delta(A)$ such that*

$$\begin{aligned} u_i(\alpha_i, a_{-i}) &\geq u_i(a_i, a_{-i}) \text{ for all } a_{-i} \in A, \text{ and} \\ u_i(\alpha_i, a_{-i}) &> u_i(a_i, a_{-i}) \text{ for some } a_{-i} \in A. \end{aligned}$$

Let $D_i(u_i) \subseteq A$ denote the set of all actions that remain after deletion of all weakly dominated actions in a game in which player i's utility function is u_i.

A game $\mathbf{u}$ *is weak dominance solvable*[9] *in one round if for every player $i \in \{R, C\}$ and for*

[9]There are various notions of "dominance solvability" in the literature. We follow Moulin (1979) except that we allow for actions to be weakly dominated by a mixed action. Note that the definition does not necessarily

all $a_i, a_i' \in D_i(u_i)$,

$$u_i(a_i, a_{-i}) = u_i(a_i', a_{-i}) \text{ for all } a_{-i} \in D_{-i}(u_{-i}).$$

Denote by **WDS** *the class of one-round weak dominance solvable games.*

Next we define a "richness" property of classes of games.

Definition 5 (Product Class) *A class of games* $\mathbf{U} \subseteq [0,1]^{2|A^2|}$ *is a product class if* $(u_i, u_{-i}), (\tilde{u}_i, \tilde{u}_{-i}) \in \mathbf{U}$ *implies* $(u_i, \tilde{u}_{-i}), (\tilde{u}_i, u_{-i}) \in \mathbf{U}$.

Product classes of games are natural to consider in the context of uncoupled learning. Uncoupledness means that the learning heuristic cannot directly depend on opponents' payoffs. If a class of games is not a product class then the opponent's payoffs are not independent from the player's payoffs. Thus, an uncoupled learning heuristic could nevertheless condition implicitly on a non-trivial subset of opponents' payoffs. Product classes of games are also motivated by our desire to find "maximal" classes of games for which we can obtain a positive result. A product class of games is closed under permutations of each player's payoff functions. Note that the class of one-round weak dominance solvable games is by definition a product class of games.

Proposition 3 *If* $\mathbf{U}$ *is a nonempty measurable class of games such that* $\mathbf{U} \subseteq$ **WDS**, *then there exists a learning heuristic that is uncoupled, converges to Nash equilibrium in all games in* $\mathbf{U}$, *and is a Nash equilibrium learning heuristic of the learning game based on* $\mathbf{U}$. *Conversely, if there is a measurable product class of games* $\mathbf{U}$ *with a pure Nash equilibrium such that* **WDS** $\subseteq \mathbf{U}$ *and for which there exists a learning heuristic that is uncoupled, converges to Nash equilibrium in all games in* $\mathbf{U}$, *and is a Nash equilibrium learning heuristic of the learning game based on* $\mathbf{U}$, *then* $\mathbf{U} \setminus$ **WDS** *must be of measure zero.*

The proof is contained in the Appendix. The first part is straightforward. The proof of the converse is by contradiction. The idea is as follows: For any generic $\mathbf{u} = (u_i, u_{-i}) \notin$ **WDS** with $\mathbf{u} \in \mathbf{U}$, we can find a generic game $\tilde{\mathbf{u}} = (\tilde{u}_i, \tilde{u}_{-i}) \in$ **WDS** such that $(u_i, \tilde{u}_{-i}) \in \mathbf{U}$ and in which player $-i$ has an incentive to pretend being in $\mathbf{u}$ although the game might be $(u_i, \tilde{u}_{-i})$. Player i behaves with uncoupled learning like in $\mathbf{u}$ giving $-i$ a higher payoff than in the unique Nash of $(u_i, \tilde{u}_{-i})$. Since these games are generic, they show up in the average long run payoff, which proves the converse.

---

imply that payoff functions are constant on *all* outcomes that remain after one round of elimination of weakly dominated actions for each player but just on outcomes that the player can unilaterally choose and that survive one round of elimination of weakly dominated actions.

There are other classes of games for which we can obtain a possibility result. In any such a class a player's own payoff is sufficiently informative about a *particular* Nash equilibrium of the game as long as one restricts to the class. One such class is the class of common interest games studied in Aumann and Sorin (1998):

Definition 6 *A game* $\mathbf{u}$ *is a* common interest *game if there is a profile of payoffs* $(z_R, z_C) \in u_R(A) \times u_C(A)$ *that strictly Pareto dominates all other profiles of payoffs, i.e.,*

$$z_R > w_R \text{ and } z_C > w_C$$

for all $(w_R, w_C) \in u_R(A) \times u_C(A) \setminus \{(z_R, z_C)\}$*, where* $u_i(A)$ *denotes the image of* u_i*.*

Let $\mathbf{CI}$ *denote the class of common interest games.*

Note that (z_R, z_C) is unique in a common interest game. Yet, action profiles for which the payoff is (z_R, z_C) may not be unique. Every common interest game has at least one pure Nash equilibrium that yields the payoff profile (z_R, z_C). Note that the class of common interest games is in some sense less "strategically trivial" than one-round weak dominance solvable games as they may still involve coordination problems. Finally, note that there one-round dominance solvable games that are not common interest games. Moreover, there are common interest games that are not one-round dominance solvable games.

Proposition 4 *If* $\mathbf{U}$ *is a nonempty measurable class of games such that* $\mathbf{U} \subseteq \mathbf{CI}$*, then there exists a learning heuristic that is uncoupled, converges to Nash equilibrium in all games* $\mathbf{U}$*, and is a Nash equilibrium learning heuristic in the learning game based on* $\mathbf{U}$*. If* $\mathbf{U}$ *is a measurable product class of games with pure Nash equilibrium such that* $\mathbf{CI} \subseteq \mathbf{U}$*, then there is no learning heuristic that is uncoupled, converges to Nash equilibrium in all games in* $\mathbf{U}$*, and is a Nash equilibrium learning heuristic in the learning game based on* $\mathbf{U}$*.*

The proof is contained in Appendix. In a common interest game, if a player obtains his maximal stage-game payoff, then she has no incentive to reach another outcome through strategic teaching. For the first part of Proposition 4 we just need to show that there is an uncoupled learning heuristics leading to a Pareto efficient Nash equilibrium in every common interest game. This is done by modifying an uncoupled learning heuristic that has been used by Hart and Mas-Colell (2006) to show convergence of uncoupled learning in games with a pure Nash equilibrium such that it now converges to efficient pure Nash equilibrium. The second part of Proposition 4 we show with a counterexample that is constructed with the help of product classes of games. Note that in contrast to the class of one-round dominance solvable games, the class of common interest games is *not* a product class. (Thus, in the second part of Proposition 4 we could have written without loss of generality $\mathbf{CI} \subsetneq \mathbf{U}$.) A player can infer from his payoff function and the fact that he faces only common interest games enough information

about the opponent's payoffs so as to know a Nash equilibrium of the game. Since the second part of Proposition 4 holds for any product class of games with pure Nash equilibrium that contains common interest games, it holds also for the smallest such class that is generated by considering all permutations over all payoff functions in common interest games. In this sense, the second part of Proposition 4 shows an impossibility even if for each player we only consider payoff functions that are consistent with some common interest game.

In some sense, both one-round dominance solvable games and common interest games are "strategically trivial" as each player can deduce from her own payoffs some Nash equilibrium action. Thus, even though both Propositions 3 and 4 are phrased as possibility results, at the heart of the matter they are impossibility results as possibilities fail when games beyond "strategically trivial" games are considered.

7 Uncoupled Strategic Teaching

We now restrict the strategic teaching heuristics to be uncoupled as well. The strategic teacher may now first learn about opponent's payoffs from opponent's behavior and then use this information to strategically teach the learning opponent. This can only work if after having learned about opponent's payoffs from opponent's behavior, the opponent's learning heuristic stays sensitive to learning so that it can be strategically taught. This is implied by uncoupledness together with finite recall and stationarity.

Definition 7 (Stationary learning heuristic with finite recall) *A learning heuristic σ_i of player i has* finite recall *if for every $\mathbf{u} \in [0,1]^{2|A^2|}$ there exists a positive integer r such that for each $t > r$, $\sigma_i^t(\mathbf{u})$ is of the form $\sigma_i^t(\mathbf{u})(\mathbf{a}^{t-r}, \mathbf{a}^{t-r+1}, ..., \mathbf{a}^{t-1})$. $\sigma_i(\mathbf{u})$ with r-recall is* stationary *if for any $t, t' > r$, $(\mathbf{a}^{t-r}, \mathbf{a}^{t-r+1}, ..., \mathbf{a}^{t-1}) = (\mathbf{a}^{t'-r}, \mathbf{a}^{t'-r+1}, ..., \mathbf{a}^{t'-1})$ implies $\sigma_i^t(\mathbf{u})(\mathbf{a}^{t-r}, \mathbf{a}^{t-r+1}, ..., \mathbf{a}^{t-1}) = \sigma_i^{t'}(\mathbf{u})(\mathbf{a}^{t'-r}, \mathbf{a}^{t'-r+1}, ..., \mathbf{a}^{t'-1})$.*

A learning heuristic satisfies finite recall if just a finite number of last periods matter instead of entire histories. It is stationarity if calendar time does not matter. Many uncoupled learning heuristics converging to Nash equilibrium used in the literature satisfy finite recall and stationarity (e.g., Hart and Mas-Colell, 2006).

To see how uncoupled convergent Nash equilibrium learning heuristics can be strategically taught by an uncoupled heuristics, consider again our example in Section 3. Assume that game $\mathbf{u}^2$ is played but neither player knows the opponent's payoff. For the first T periods, let both players follow an uncoupled learning heuristic converging to Nash equilibrium in all games. As T becomes larger and larger, the probability of being away from playing Nash equilibrium (b, b) should become small. Now suppose that after some finite period T, Rowena switches to a learning heuristic that behaves in game $\mathbf{u}^2$ as if playing game $\mathbf{u}^1$. If Colin's learning heuristics

has finite recall, is uncoupled, and Nash equilibrium convergent, play (re-)converges almost surely to (c, a) in game $\mathbf{u}^2$. This yields her a strictly higher long-run payoff than with the learning heuristic converging to the Nash equilibrium in $\mathbf{u}^2$. That is, as T increases, Rowena is able to use the first T periods to learn about Colin's best response in $\mathbf{u}^2$ and then switches heuristics so as if she learns in $\mathbf{u}^1$ thus effectively misleading Colin to "think" that game $\mathbf{u}^1$ is being played. Note that Colin cannot deduce from Rowena's play that they are not playing $\mathbf{u}^1$ since his learning heuristics is uncoupled and has finite recall. While this argument works when Rowena's teaching heuristic be uncoupled, it relies on Rowena having a sufficiently larger memory than Colin. The proof of the following observation now follows from above arguments and arguments made in Section 3.

Proposition 5 *There is no stationary uncoupled learning heuristic with finite recall that both converges to Nash equilibrium in all finite games and is a Nash equilibrium learning heuristic of the learning game when restricting to the set of heuristics Σ to uncoupled heuristics only.*

As with Proposition 1, the observation continues to hold even if we restrict $\mathbf{U}$ to be the class of all generic finite games, all two-player games, all 2×2 games, all games with pure Nash equilibrium, all games with a unique Nash equilibrium, all games with strategic complements, all ordinal potential games, or any union thereof, or if we weaken convergence to correlated equilibrium, the set of iterated admissible strategy profiles, rationalizable strategy profiles, or minimal curb sets.

Proposition 2 on the incentives for strategic teaching can also be extended to uncoupled strategic teaching heuristics if the opponent uses a stationary uncoupled learning heuristic with perfect recall.

Proposition 6 *Let $\mathbf{U}$ be the class of finite two-player games that possess a pure Nash equilibrium. For any profile of stationary uncoupled learning heuristics with finite recall (σ_R, σ_C) that converges to Nash equilibrium in all games in $\mathbf{U}$, there exists an uncoupled learning heuristic $\tilde{\sigma}$ (i.e., an uncoupled "strategic teaching heuristic") such that when Rowena follows $\tilde{\sigma}$ and Colin follows σ_C, we have*

$$V_R((\tilde{\sigma}, \sigma_C), \mathbf{U}) \geq L_R(\mathbf{U}) > V_R((\sigma_R, \sigma_C), \mathbf{U}).$$

To prove the result, let U_i denote the projection of U on the i-coordinate. Since $A \times A$ is finite, there exists a finite partition of U_i such that i's pure best response structure is identical within each partition cell. We only need to focus on generic games. Thus, discard all partition cells with Lebesque measure zero (i.e., with indifferences). Since the number of (remaining) partition cells are finite, enumerate them $1, ..., n$. Let Rowena's uncoupled strategic teaching heuristic be such that she follows an uncoupled learning heuristic converging to Nash equilibrium in all games for the first nT periods, resetting it after every T periods. Let Rowena pretend

to play a game with pure best response structure of partition cell 1 in the first T periods, a game with pure best response structure of partition cell 2 in the next periods $T+1, ..., 2T$, and let her continue in this fashion to the game with pure best response of partition cell n in the last $(n-1)T+1, ..., nT$ periods. For generic games, one of the partition cells corresponds to Rowena's actual ordinal payoff structure. As T becomes larger and larger, the probability of being away from playing a pure Nash equilibrium in each of the n games becomes small. After nT periods, Rowena learned all relevant best responses of Colin. She can now pretend to play the game with the pure equilibrium corresponding to the Stackelberg outcome of the actual game. Such game exists as constructed in the proof of Proposition 2. Colin eventually plays the Stackelberg follower action of the actual game since he uses a stationary uncoupled learning heuristics with finite recall converging to Nash equilibrium of the "pretended" game. The rest of the proof now follows from arguments in the proof of Proposition 2.

The "positive" results of Propositions 3 and 4 can also be extended to uncoupled strategic teaching heuristics when the opponent uses a stationary uncoupled learning heuristic with finite recall. First, it is trivial that stage-game equilibrium in dominance solvable games can be reached with a stationary uncoupled learning heuristic with finite recall converging to stage-game equilibrium. Moreover, the learning heuristics constructed in the proof of Proposition 4 for converging to Pareto dominant equilibrium is stationary and has finite recall. The converses follow from arguments analogous to just presented to prove Proposition 6 and the proofs of Propositions 3 and 4.

8 Further Extensions

8.1 Discounting

So far, we defined learning games only with limit of means payoffs. Our results remain valid with other forms of time preferences as long as players are sufficiently patient. Consider the expected discounted payoffs defined for player $i \in \{R, C\}$ by

$$v_i^\delta(\boldsymbol{\sigma}(\mathbf{u})) := \mathbb{E}_{\boldsymbol{\sigma}(\mathbf{u})}\left[(1-\delta)\sum_{t=1}^{\infty}\delta^{t-1}u_i(\mathbf{a}^t(\boldsymbol{\sigma}(\mathbf{u})))\right], \tag{3}$$

for $\delta \in [0, 1)$. Given δ, the long-run expected payoffs "averaged" over all games in the Lebesgue measurable subset of games $\mathbf{U} \subseteq [0,1]^{2|A^2|}$ is

$$V_i^\delta(\boldsymbol{\sigma}, \mathbf{U}) := \int_{\mathbf{U}} v_i^\delta(\boldsymbol{\sigma}(\mathbf{u}))d\lambda. \tag{4}$$

We now consider the δ-discounted learning game $\langle\{R, C\}, \Sigma, (V_i^\delta(\cdot, \mathbf{U}))_{i=R,C}\rangle$ parameterized by the common discount factor $\delta \in [0, 1)$.

A profile of learning heuristics $\boldsymbol{\sigma} = (\sigma_R, \sigma_C) \in \Sigma \times \Sigma$ is a Nash equilibrium of the δ-discounted learning game $\langle \{R, C\}, \Sigma, (V_i^\delta(\cdot, \mathbf{U}))_{i=R,C} \rangle$ if for all $i \in \{R, C\}$,

$$V_i^\delta(\sigma_i, \sigma_{-i}, \mathbf{U}) \geq V_i^\delta(\hat{\sigma}_i, \sigma_{-i}, \mathbf{U}) \text{ for all } \hat{\sigma}_i \in \Sigma. \tag{5}$$

Our observation holds now for learning games in the limit as $\delta \to 1$.

Proposition 7 *In the limit, as $\delta \to 1$ there is no uncoupled learning heuristic that both converges to Nash equilibrium in all finite games and is a Nash equilibrium learning heuristic of the δ-discounted learning game.*

The proof follows from Proposition 1 and the Hardy-Littlewood Theorem. Note that in the proof of Proposition 1, no matter whether Rowena uses the strategic teaching heuristic or a Nash equilibrium learning heuristic, the profiles of heuristics are convergent in the games considered in the counterexample since we assume that the learning heuristics converge to Nash equilibrium in every finite game with pure Nash equilibrium. For convergent sequences, the lim inf of means payoffs is equal to the lim sup of means payoffs. The Hardy-Littlewood Theorem (e.g., Maschler, Solan, and Zamir, 2013, Theorem 13.31) implies that the limit of δ-discounted payoffs is "sandwiched" between the lim inf of means payoff and lim sup of means payoffs when δ goes to 1. Thus, for the sequences considered for the proof, the limit of the δ-discounted payoffs must be equal to the lim inf of means payoffs as δ goes to 1. The rest of the proof now follows from the proof of Proposition 1.

8.2 $(1-\varepsilon)$-Convergence and Approximate Equilibrium

Various notions of convergence have been used in the literature on learning in games. Some of the literature uses convergence to ε-Nash equilibrium $(1-\varepsilon)$ of the time after some sufficiently long time period (e.g., Foster and Young, 2003, 2006). Compared to Section 2, this is a weaker notion of convergence and a weaker equilibrium notion. An ε-Nash equilibrium is a profile of mixed actions for which neither player can increase its payoff by more than ε through a unilateral change of actions. We generalize our observation in Section 3 to $(1-\varepsilon)$-convergence and ε-Nash equilibrium.

Proposition 8 *There is an $\bar{\varepsilon} > 0$ such that for any $\varepsilon \in [0, \bar{\varepsilon}]$, there is no learning heuristic that is uncoupled, converges in every game to ε-Nash equilibrium with probability of at least $(1-\varepsilon)$ as t becomes large, and is a Nash equilibrium of the learning game.*

The proof in Appendix generalizes our counterexample by showing that there exists an $\bar{\varepsilon} > 0$ such for every $\varepsilon \in [0, \bar{\varepsilon}]$ we can find payoffs for a counterexample similar to the one in Section 3. In both games of the counterexample, the unique pure Nash equilibrium is also the unique ε-Nash equilibrium. Moreover, games are generic.

8.3 Mixed Equilibrium and Strictly Competitive Games

We considered learning heuristics that converge to pure Nash equilibrium of the stage-game if such an equilibrium exists. Are our observations an artefact of focusing on pure equilibrium? Would it be possible to find uncoupled learning heuristics that converge to non-degenerate mixed equilibrium of the stage-game (in games that do have such equilibria) and that each player has an incentive to adopt if the opponent adopt it as well? Moreover, is there a possibility result when restricting to strictly competitive games, a class of games in which learning is usually "nice"? Note that when confining the analysis to strictly competitive games, the uncoupledness assumption of learning heuristics loses to a large extent its restrictiveness since each player's payoff function is informative about the opponent's payoff function.

Consider as a counterexample the "matching pennies" game

	h	t
h	$1,-1$	$-1,1$
t	$-1,1$	$1,-1$

The unique Nash equilibrium is the profile of non-degenerate mixed actions, $((\frac{1}{2},\frac{1}{2}),(\frac{1}{2},\frac{1}{2}))$. Now consider a "biased matching pennies" game,

	h	t
h	$2,-1$	$-1,1$
t	$-1,1$	$1,-1$

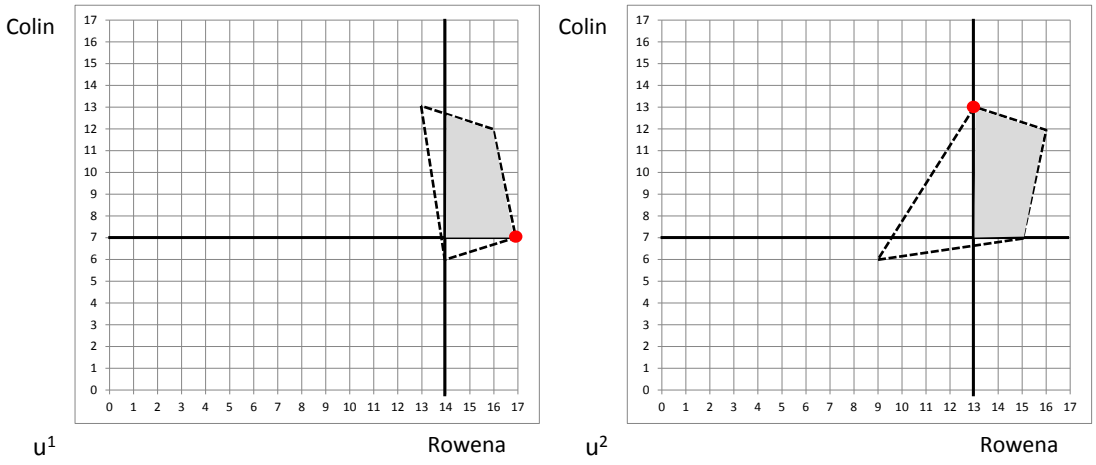
that just differs from the previous game in that Rowena's payoff from (h,h) increased to 2. Colin's payoffs remain unchanged. Since Rowena's equilibrium mixed action must make Colin indifferent among his actions, Rowena's equilibrium mixed action remains unchanged. Colin's equilibrium action changes to $(\frac{2}{5},\frac{3}{5})$. He should put less weight on pure action h so as to keep Rowena indifferent among her actions. Since Colin uses an uncoupled learning heuristic, his behavior in both games must be the same unless Rowena "communicates" her change of payoff through her different play. Note, however, that Rowena does not have an incentive to do so. If she "communicates" her biased payoff and both players use learning heuristics leading to Nash equilibrium (let's say in terms of almost-sure convergence of per-period behavior à la Hart and Mas-Colell, 2006, Theorem 7) in both games, then her long run payoff is 0.2. If instead she behaves in the biased matching pennies game like her learning heuristic in the matching pennies game, then Colin is not able to learn about her biased payoffs and her long run payoff is 0.25. Since both games are generic, this simple example shows that we can apply arguments similar to the previous sections also to the case of convergence to non-degenerate mixed equilibrium and the class of strictly competitive games.

9 Discussion

Inventive Compatibility and Repeated Games with Incomplete Information

Uncoupled learning heuristics converging to stage-game equilibrium implicitly "communicate" through behavior information about the player's payoff to other players. This is how those profiles of learning heuristics eventually find stage-game Nash equilibrium. Our observations suggest that such an implicit communication of payoffs through learning heuristics is not long-run incentive compatible in all games. This is reminiscent of repeated games with incomplete information. To see this, consider our counterexample in Section 3 as a repeated game with one-sided incomplete information and known own payoff matrices. At the beginning of the repeated game, either payoff matrix $\mathbf{u}^1$ or $\mathbf{u}^2$ is drawn according to some non-degenerate probability distribution. Rowena is informed about the draw while Colin is kept ignorant. Note that Colin's payoff matrix in game $\mathbf{u}^1$ is identical to his payoff matrix in game $\mathbf{u}^2$. Thus, he has complete information about his own payoff matrix but incomplete information about his opponent's payoff matrix. According to a characterization result by Shalev (1994), every Bayesian Nash equilibrium of this repeated game is payoff-equivalent to a fully revealing Nash equilibrium of the repeated game. Figure 1 shows the payoffs for the repeated games $\mathbf{u}^1$ (left figure) and $\mathbf{u}^2$ (right figure). The area shaped by the intermitted line indicates feasible payoffs of the repeated

Figure 1: Nash Equilibrium Payoffs of the Repeated Games in the Counterexample



game with complete information. The thick vertical and horizontal solid lines mark the minimax payoffs of Rowena and Colin in those games, respectively. Consequently, the grey-shaded areas show equilibrium payoffs in the repeated game with complete information. The red dot indicates the stage-game Nash equilibrium payoff. There is no Bayesian Nash equilibrium of the repeated game with one-sided incomplete information in which players play like the repeated stage-game Nash equilibrium in $\mathbf{u}^1$ when $\mathbf{u}^1$ is drawn and the repeated stage-game Nash equilibrium in $\mathbf{u}^2$ when $\mathbf{u}^2$ is drawn. A necessary condition of Shalev (1994)'s characterization result is *incentive*

compatibility for Rowena. In our setting this means that Rowena's payoff from playing $\mathbf{u}^2$ must be weakly larger than her payoff in $\mathbf{u}^2$ when playing like in $\mathbf{u}^1$. Clearly, this is violated in our counterexample as she receives 13 in the stage-game Nash equilibrium of $\mathbf{u}^2$ but 15 when she plays in $\mathbf{u}^2$ like in (the stage-game Nash equilibrium of) $\mathbf{u}^1$.

In our counterexample it is not incentive compatible for Rowena to implicitly communicate here payoffs to Colin via the equilibrium learning heuristic. She has an incentive to lie about her payoffs and strategically teach Colin her "wrong" payoffs. Her strategic teaching strategy of playing in $\mathbf{u}^2$ like in $\mathbf{u}^1$ is the *best* strategy against Colin. This follows from a characterization result by Israeli (1999) in the repeated games literature. Applied to our context, his result means that the best Rowena can do against Colin in $\mathbf{u}^2$ is to play as if she faces a zero-sum game in which her payoffs are the negative of Colin's payoffs. Note that this zero-sum game has the same pure best response structure as $\mathbf{u}^1$. Rowena's stage-game Nash equilibrium strategy in $\mathbf{u}^1$ minimizes Colin's payoff in $\mathbf{u}^2$. Thus, if she plays $\mathbf{u}^2$ like $\mathbf{u}^1$, she plays like in the zero-sum game.

Population Games

One reviewer pointed out that strategic considerations at the core of our counterexample are mostly absent when learning takes place in population games. Indeed in a review paper, Fudenberg and Levine (1998b) remark that "(m)ost of learning theory abstracts from these repeated game considerations by explicitly or implicitly relying on a model in which the incentive to try to alter the future play of opponents is small enough to be negligible. This can be justified by appeal to models with a large number of players, who interact anonymously (which is the case in most experiments), with the population size large compared to the discount factor." Yet, they also pointed to Ellsion (1997) as a caveat. Ellison (1997, Section 5) showed in an interesting example of a 3×3 game that strategic teaching of fictitious play with finite memory is possible even in a population game context due to contagion, and that strategic teaching may become more effective the larger the population size. Thus, large populations are neither sufficient nor necessary for rendering strategic teaching mute.

We do not know yet how to extend our observations to a setting in with a population of Rowenas and Colins, or a setting in which the strategic teacher enters a population of learners who are randomly matched to play a game in various positions.[10] Instead we focus on games without a population content. There are at least two justifications for this. First, in games without population context, it is not immediate that strategic teaching must necessarily interfere with learning stage-game equilibrium as the repeated stage-game equilibrium is also an equilibrium of the repeated game. What we show is that uncoupledness of the learning

[10]It is immediate though to extend our observations to one Rowena playing against a population of Colins. Then the same arguments apply although convergence may be slower.

heuristics creates this interferences. Second, we are line with recent papers on uncoupled equilibrium learning who do not make use of population games (e.g., Babichenko, 2010, Foster and Young, 2001, 2003, 2006, Hart and Mas-Colell, 2003, 2006, Germano and Lugosi, 2007, Kakade and Foster, 2008, Young, 2009). It is perhaps not surprising that this literature does not assume population games as they seek a theory of learning equilibrium in general strategic games and not just in population games.

Experimental Evidence for Strategic Teaching

That uncoupled equilibrium learning gives rise to opportunities for strategic teaching is not just a theoretical curiosity. There is experimental evidence (Duersch et al, 2010, Terracol and Vaksmann, 2009, Camerer et al., 2002, Chong et al., 2006, Hyndman et al., 2012) that some participants do indeed use sophisticated behavior akin to strategic teaching when opponents follow uncoupled learning heuristics. E.g., Duersch et al. (2010) present stark example of strategic teaching in an experiment in which human subjects played against a computer opponent programmed to various learning heuristics. Terracol and Vaksmann (2009) show in an experimental 3×3 game that players may forego some immediate payoff in order to modify the opponent's future behavior.

Evolutionary Stability of Learning Heuristics

Our observations can be interpreted in an *evolutionary* context. Instead of players choosing strategically learning heuristics with the long-run benefit in mind, the emergence of learning heuristics may be the result of mutations and evolutionary selection. Suppose there is a large population of players who are randomly matched to play two-player games drawn at random from a class of games $\mathbf{U}$. For each game, each player has equal chance to play in the position of the row or column player (which symmetrizes games). Players are programmed to learning heuristics. Each player's fitness (e.g., offsprings) is measured by the long-run payoffs "averaged" over those games. (That is, evolution is assumed to be slower than it takes for average long-run payoffs from learning to emerge.) Suppose now that initially the population is programmed to an uncoupled learning heuristic converging to Nash equilibrium in all games. Would such a population be robust to a small fraction of mutants that may invade with another heuristic to play in those games? Although the notion of evolutionary stability is intricate in repeated games (see for instance Binmore and Samuelson, 1992, Demichelis, 2013, Fudenberg and Maskin, 1990, Kim, 1994), we believe that a reasonable notion of evolutionary stability in repeated games would require Nash equilibrium of the repeated game as a necessary condition. Under this assumption, we conclude from our observations that if an uncoupled learning heuristic converges to Nash equilibrium in all games, then it cannot be evolutionary stable.

Learning to Learn

A third interpretation of our observations is based on the idea that players may also learn how to learn in games and the quest for universal learning heuristics. Since learning heuristics apply to all games, we can apply them also to the learning game. Our observations suggest that there is no uncoupled learning heuristic converging to Nash equilibrium in all finite games that could learn itself.[11] This is in contrast to the existence of universal learning heuristics for "single-person decision problems" in artificial intelligence (Schmidhuber, 2003).

Related Literature

Our paper contributes to the literature discussing the limits of strategic learning (Foster and Young, 2001, Hart and Mas-Colell, 2003, 2006, Jordan 1993, Nachbar 1997, 2001, 2005, Sadzik, 2011). For instance, Nachbar (1997, 2001, 2005) pointed out a tension between prediction and optimization in rational learning. Kalai and Lehrer (1993) studied rational learners who form suitable probabilistic beliefs about opponents' strategies in infinitely repeated games, update beliefs according to Bayes rule, and chose strategies so as to maximize discounted expected utility. They show that rational learners must eventually play like a Nash equilibrium of the *repeated game*. Nachbar (1997, 2001, 2005) showed that since the set of possible strategies is large in repeated games, there is no belief that allows a rational learner to predict future play for every possible strategy of the opponent. In some sense, for rational learning to work, prior beliefs must already be in a kind of "pre-equilibrium". Another problem was studied by Sadzik (2011). He focuses on uncoupled learning heuristics that converge to equilibrium. He shows that finding among uncoupled equilibrium learning heuristics a pair that actually converges requires a large degree of ex-ante coordination on learning heuristics. While these two impossibility results are clearly different from our observations, all have in common the implicit question of how would players come up with "right" learning strategies or beliefs.

The theoretical literature on strategic teaching of learning players is small. Previous papers focused on special classes of games and particular learning heuristics while we seek general results. Fudenberg and Kreps (1993) pointed out with an example the possibility of Stackelberg leadership against an opponent following fictitious play. Ellison (1997) shows that a single rational player in a population of players learning by fictitious play can strategically teach the selection of risk dominant equilibria in 2×2 coordination games. He also presents simulation results for nonmyopic manipulation in a 3×3 game. Schipper (2019) derives optimal strategic

[11] If there were, we would have pushed the problem just one level further. How do players were to find such a learning heuristic? Presumably there is a (meta-meta-)learning heuristic to (meta-)learn the learning heuristic to play in the games ... Clearly, we are headed for an infinite regress akin to "how to decide how to decide ..." that have been studied by Mongin and Walliser (1988), Lipman (1991), and Ergin and Sarver (2010). There is no need for us to formalize this infinite regress problem here because our counterexample shows a problem already at the second level.

teaching strategies against an opponent learning according to myopic best response in games with strategic complements or substitutes. Kordonis et al. (2018) show the possibility of Stackelberg leadership against some types of adaptive players in games motivated by electricity markets. Duersch et al. (2012, 2014) and Schipper (2009) characterize the class of games in which conditional and unconditional imitation strategies can not be taken advantage of.

A Proofs

A.1 Proof of Proposition 2

Fix a generic stage-game $\mathbf{u} = (u_i, u_{-i}) \in \mathbf{U}$. In generic games best responses are unique. Hence, we write

$$a_i^L := \arg\max_{a_i \in A} u_i(a_i, a_{-i}) \text{ s.t. } a_{-i} = B_{-i}(\mathbf{u})(a_i)$$

and

$$a_{-i}^F := B_{-i}(\mathbf{u})(a_i^L)$$

for the Stackelberg leader i and follower $-i$'s actions of the stage-game $\mathbf{u}$, respectively.

Since $\mathbf{U}$ is the class of finite two-player games that possess a pure Nash equilibrium, there exists a game $(\tilde{u}_i, u_{-i}) \in \mathbf{U}$ such that

(a) (a_i^L, a_{-i}^F) is the unique Nash equilibrium of $(\tilde{u}_i, u_{-i})$, and

(b) $\tilde{u}_i(a_i^L, a_{-i}^F) \geq \ell_i(\tilde{u}_i, u_{-i})$.

E.g., let $\tilde{u}_i$ be such that

(A) a_i^L strictly dominates all other actions in A for player i, i.e., for any $\alpha_i \in \Delta(A)$ with $\alpha_i(a_i^L) < 1$, we have $\tilde{u}_i(a_i^L, a_{-i}) > \tilde{u}_i(\alpha_i, a_{-i})$ for any $a_{-i} \in A$,

(B) player i's payoff from the profile (a_i^L, a_{-i}^F) strictly dominates any other payoff, i.e., $\tilde{u}_i(a_i^L, a_{-i}^F) > \tilde{u}_i(\mathbf{a})$ for any $\mathbf{a} \in A \times A$, $\mathbf{a} \neq (a_i^L, a_{-i}^F)$.

(A) implies (a) since $(\tilde{u}_i, u_{-i})$ is generic, a_i^L is the unique best response of player i to a_{-i}^F in $(\tilde{u}_i, u_{-i})$ and $a_{-i}^F = B_{-i}(\mathbf{u})(a_i^L) = B_{-i}(\tilde{u}_i, u_{-i})(a_i^L)$. (B) implies (b). Moreover, both (A) and (B) are consistent in the sense that for any generic $\mathbf{u} \in \mathbf{U}$ there exist $(\tilde{u}_i, u_{-i}) \in \mathbf{U}$ such that both (A) and (B) hold.

Since (σ_i, σ_{-i}) is a profile of uncoupled learning heuristics that converges almost surely to the stage-game Nash equilibrium in all games in $\mathbf{U}$, it must almost surely converge to the Nash equilibrium (a_i^L, a_{-i}^F) in the game $(\tilde{u}_i, u_{-i})$. Let player i now follow a learning heuristic $\tilde{\sigma}_i$ that in both games, $\mathbf{u}$ and $(\tilde{u}_i, u_{-i})$, behaves like σ_i in game $(\tilde{u}_i, u_{-i})$. (Thus, when in $\mathbf{u}$, player i "pretends" to be in $(\tilde{u}_i, u_{-i})$.) Since σ_{-i} is an uncoupled learning heuristic, player $-i$ behaves

almost surely identically in $\mathbf{u}$ and $(\tilde{u}_i, u_{-i})$ when player i follows $\tilde{\sigma}_i$. Thus, in both games, $\mathbf{u}$ and $(\tilde{u}_i, u_{-i})$, almost every play path consists of (a_i^L, a_{-i}^F) being played from some point on. Hence,

$$\lim_{T\to\infty} \inf \mathbb{E}_{(\tilde{\sigma}_i(\mathbf{u}),\sigma_{-i}(\mathbf{u}))} \left[\frac{1}{T} \sum_{t=1}^{T} u_i(\mathbf{a}^t(\tilde{\sigma}_i(\mathbf{u}), \sigma_{-i}(\mathbf{u}))) \right] \geq \ell_i(\mathbf{u})$$

and

$$\lim_{T\to\infty} \inf \mathbb{E}_{(\tilde{\sigma}_i(\tilde{u}_i,u_{-i}),\sigma_{-i}(\tilde{u}_i,u_{-i}))} \left[\frac{1}{T} \sum_{t=1}^{T} \tilde{u}_i(\mathbf{a}^t(\tilde{\sigma}_i(\tilde{u}_i, u_{-i}), \sigma_{-i}(\tilde{u}_i, u_{-i}))) \right] \geq \ell_i(\tilde{u}_i, u_{-i}).$$

Since the argument holds for almost any repeated stage-game $\mathbf{u} \in \mathbf{U}$, we must have

$$V_i((\tilde{\sigma}_i, \sigma_{-i}), \mathbf{U}) \geq L_i(\mathbf{U}).$$

It is known that there are finite two-player games with a pure Nash equilibrium in which the (worst) Stackelberg leader payoff as defined here can be strictly below a Nash equilibrium payoff (e.g., Başar and Olsder, 1999, p. 132-133). But such games must be non-generic. In a generic game, best responses are unique. Thus, in generic games a Stackelberg leader can guarantee herself at least her best Nash equilibrium payoff because if the Stackelberg leader chooses the action corresponding to her most preferred pure Nash equilibrium, then the opponent best responds uniquely with his corresponding Nash equilibrium action. When "averaging" limit-expected-mean payoffs over all games in $\mathbf{U}$ with a Lebesgue measure, non-generic games must have measure zero. Thus $L_i(\mathbf{U}) \geq V_i((\sigma_i, \sigma_{-i}), \mathbf{U})$.

In our example $\mathbf{u}^2$ from Section 3 we observe that the Stackelberg leader achieves a payoff that is strictly larger than in the unique Nash equilibrium. Since $\mathbf{u}^2$ is generic, this holds for an open neighborhood $\mathbf{U}^2 \subseteq \mathbf{U}$ of the stage-game. Since any nonempty open neighborhood must have strict positive Lebesgue measure, we must have $L_i(\mathbf{U}) > V_i((\sigma_i, \sigma_{-i}), \mathbf{U})$.

Finally, note that in above arguments, (a_i^L, a_{-i}^F) is the unique and strict Nash equilibrium of $(\tilde{u}_i, u_{-i})$ in which player i plays a strict dominant action. Thus, it is also the correlated equilibrium, iterative admissible action profile, rationalizable action profile, and minimal curb set of $(\tilde{u}_i, u_{-i})$. Hence, the result holds also for uncoupled learning heuristics that converge to correlated equilibrium, iterative admissible action profiles, rationalizable action profiles or minimal curb sets, respectively. This completes the proof of the proposition. □

A.2 Proof of Proposition 3

Let $\mathbf{U} \subseteq \mathbf{WDS}$ with $\mathbf{u} \in \mathbf{U}$. By definition of **WDS**, any action that remains after one round of elimination of weakly dominated actions in the stage-game $\mathbf{u}$ is a Nash equilibrium action. For each player i, select σ_i that in every stage-game $\mathbf{u} \in \mathbf{U}$ chooses an action $a_i \in D_i(u_i)$ for every history. Then the profile $\boldsymbol{\sigma} = (\sigma_i, \sigma_{-i})$ selects a Nash equilibrium in every stage-game

$\mathbf{u} \in \mathbf{U}$. Require also that the action that σ_i selects for player i in $(u_i, \tilde{u}_{-i}) \in \mathbf{U}$ is identical to the action that it selects in $\mathbf{u} \in \mathbf{U}$, for every $\tilde{u}_{-i}$. Then σ_i is uncoupled. Moreover, it follows that the profile of such learning heuristics is a Nash equilibrium in the learning game based on $\mathbf{U}$.

We prove the converse by contradiction. Let $\mathbf{WDS} \subseteq \mathbf{U}$ and let $\mathbf{u} \in \mathbf{U}$ be a generic stage-game with $\mathbf{u} \notin \mathbf{WDS}$. Fix a history of play of $\mathbf{u}$ that emerges from the profile of uncoupled learning heuristics $\boldsymbol{\sigma} = (\sigma_i, \sigma_{-i})$ that converges to Nash equilibrium in all games in $\boldsymbol{U}$. Since every game in $\boldsymbol{U}$ has a pure Nash equilibrium by definition, there is a Nash equilibrium of the stage-game $\mathbf{u}$ to which $\boldsymbol{\sigma}(\mathbf{u})$ converges almost surely. Denote it by $\boldsymbol{a}^\infty(\boldsymbol{\sigma}(\mathbf{u})) = (a_i^\infty(\boldsymbol{\sigma}(\mathbf{u})), a_{-i}^\infty(\boldsymbol{\sigma}(\mathbf{u})))$.

We claim that there exists a generic stage-game $\tilde{\mathbf{u}} = (\tilde{u}_i, \tilde{u}_{-i}) \in \mathbf{WDS}$ such that

(0) $(u_i, \tilde{u}_{-i})$ is a generic game in $\mathbf{U}$,

(i) any Nash equilibrium action of player i in the stage-game $(u_i, \tilde{u}_{-i})$ differs from $a_i^\infty(\boldsymbol{\sigma}(\mathbf{u}))$,

(ii) $\tilde{u}_{-i}(\boldsymbol{a}^\infty(\boldsymbol{\sigma}(\mathbf{u}))) > \tilde{u}_{-i}(\boldsymbol{a}^*)$ for any Nash equilibrium $\boldsymbol{a}^*$ of the stage-game $(u_i, \tilde{u}_{-i})$.

Intuitively, (ii) implies that player $-i$ has an incentive to pretend being in game $\mathbf{u}$ even though the stage-game is $(u_i, \tilde{u}_{-i})$. Together with (i), player $-i$ has an incentive to not let the play converge to a Nash equilibrium of $(u_i, \tilde{u}_{-i})$. (0) implies that this is relevant for the learning game based on $\mathbf{U}$.

(0) follows because $\tilde{\mathbf{u}} \in \mathbf{WDS} \subseteq \mathbf{U}$, both $\tilde{\mathbf{u}}$ and $\mathbf{u}$ are generic, $\mathbf{U}$ is a product class of finite two-player games, and $(u_i, \tilde{u}_{-i})$ has a pure Nash equilibrium. To see the last point, pick for player $-i$ an action $a_{-i} \in D_{-i}(\tilde{u}_{-i})$ and note that because $\tilde{\mathbf{u}}$ is generic there must be a pure and unique best response a_i to it by player i. Action a_{-i} is the unique best response to a_i in $(u_i, \tilde{u}_{-i})$ because $\tilde{\mathbf{u}} \in \mathbf{WDS}$ and $\tilde{\mathbf{u}}$ is generic. Thus, (a_i, a_{-i}) is a strict pure Nash equilibrium of $(u_i, \tilde{u}_{-i})$.

(i) follows from $\mathbf{u} \notin \mathbf{WDS}$. To see this, note that $\mathbf{u} \notin \mathbf{WDS}$ implies that for some player i there exist $a_i^{**}, a_i^* \in D_i(u_i)$ such that

$$u_i(a_i^{**}, a_{-i}) \quad > \quad u_i(a_i^*, a_{-i}) \text{ for some } a_{-i} \in D_{-i}(u_{-i}).$$

If in addition

$$u_i(a_i^{**}, a_{-i}) \quad \geq \quad u_i(a_i^*, a_{-i}) \text{ for all } a_{-i} \in A,$$

then a_i^{**} weakly dominates a_i^*, a contradiction to $a_i^* \in D_i(u_i)$. (In the last inequality, we could have written ">" since the game is generic.) Thus, we must also have

$$u_i(a_i^{**}, a_{-i}) < u_i(a_i^*, a_{-i}) \text{ for some } a_{-i} \in A. \tag{6}$$

Since $\mathbf{u}$ is generic, there exist $a_{-i}^{**}, a_{-i}^{*} \in A$ such that a_i^{**} is the unique best response to a_{-i}^{**} and a_i^{*} is the unique best response to a_{-i}^{*}.

Since $\boldsymbol{\sigma}$ converges to a pure Nash equilibrium of $\mathbf{u}$, we must have $a_i^{\infty}(\boldsymbol{\sigma}(\mathbf{u})) \neq a_i^{**}$ or $a_i^{\infty}(\boldsymbol{\sigma}(\mathbf{u})) \neq a_i^{*}$. With loss of generality, assume the latter case (otherwise replace * with ** in below arguments), $a_i^{\infty}(\boldsymbol{\sigma}(\mathbf{u})) \neq a_i^{*}$. We can choose a generic stage-game $\tilde{\mathbf{u}} \in \mathbf{WDS}$ such that a_{-i}^{*} strictly dominates all other actions in A for player $-i$, i.e., for all $a_{-i} \in A \setminus \{a_{-i}^{*}\}$,

$$\tilde{u}_{-i}(a_i, a_{-i}^{*}) > \tilde{u}_{-i}(a_i, a_{-i}) \text{ for all } a_i \in A.$$

Then by construction, a_{-i}^{*} is the unique Nash equilibrium action of player $-i$ in the game $(u_i, \tilde{u}_{-i})$. Hence, $\mathbf{a}^{*} = (a_i^{*}, a_{-i}^{*})$ is the unique Nash equilibrium of $(u_i, \tilde{u}_{-i})$. It follows that player i's Nash equilibrium action in the game $(u_i, \tilde{u}_{-i})$ is different from $a_i^{\infty}(\boldsymbol{\sigma}(\mathbf{u}))$, which finishes the proof of (i).

To prove (ii), note that by previous arguments it is sufficient to show

$$\tilde{u}_{-i}(\boldsymbol{a}^{\infty}(\boldsymbol{\sigma}(\mathbf{u}))) > \tilde{u}_{-i}(\mathbf{a}^{*}),$$

for the unique Nash equilibrium $\mathbf{a}^{*}$ of the game $(u_i, \tilde{u}_{-i})$.

Note that $a_{-i}^{\infty}(\boldsymbol{\sigma}(\mathbf{u})) \neq a_{-i}^{*}$. Suppose not, then since a_i^{*} is the unique best response to a_{-i}^{*} in the stage-game $\mathbf{u}$, it follows that $a_i^{\infty}(\boldsymbol{\sigma}(\mathbf{u})) = a_i^{*}$, a contradiction to the assumption above that $a_i^{\infty}(\boldsymbol{\sigma}(\mathbf{u})) \neq a_i^{*}$.

We can choose $\tilde{u}_{-i}$ such that

$$\tilde{u}_{-i}(a_i^{\infty}(\boldsymbol{\sigma}(\mathbf{u})), a_{-i}^{*}) = \tilde{u}_{-i}(\boldsymbol{a}^{\infty}(\boldsymbol{\sigma}(\mathbf{u}))) + \varepsilon = \tilde{u}_{-i}(\mathbf{a}^{*}) + 2\varepsilon$$

for some $\varepsilon > 0$. This makes $\tilde{u}_{-i}(\boldsymbol{a}^{\infty}(\boldsymbol{\sigma}(\mathbf{u})))$ sufficiently large in order to satisfy

$$\tilde{u}_{-i}(\boldsymbol{a}^{\infty}(\boldsymbol{\sigma}(\mathbf{u}))) \quad > \quad \tilde{u}_{-i}(\mathbf{a}^{*}),$$

while continuing to satisfy

$$\tilde{u}_{-i}(a_i^{\infty}(\boldsymbol{\sigma}(\mathbf{u})), a_{-i}^{*}) \quad > \quad \tilde{u}_{-i}(\boldsymbol{a}^{\infty}(\boldsymbol{\sigma}(\mathbf{u}))),$$

a necessary condition for a_{-i}^{*} being strict dominant. This finishes the proof of (ii). Note that (i) and (ii) (and (0)) can be satisfied simultaneously.

Note that if player i adopts σ_i, then player $-i$ can strictly improve her long run payoff in the repeated stage-game $(u_i, \tilde{u}_{-i})$ with σ_{-i}^{*} that satisfies $\sigma_{-i}^{*}(\mathbf{u}) = \sigma_{-i}^{*}(u_i, \tilde{u}_{-i}) = \sigma_{-i}(\mathbf{u})$. That is, in game $(u_i, \tilde{u}_{-i})$, player $-i$ pretends to be in $\mathbf{u}$. Since both stage-games $\mathbf{u}$ and $(u_i, \tilde{u}_{-i})$ are generic, our arguments above hold also for open neighborhoods of them. Since any nonempty open neighborhoods must have strict positive Lebesgue measure, player $-i$'s "average" long-run payoff from σ_{-i}^{*} in the learning game when player i follows σ_i is strictly larger than from σ_{-i}. Thus, $\boldsymbol{\sigma}$ is not a Nash equilibrium of the learning game, a contradiction.

In the above arguments, $(u_i, \tilde{u}_{-i})$ has a unique and strict Nash equilibrium in which one player plays a strict dominant action. Thus, it must also be the correlated equilibrium, the iterative admissible action profile, the rationalizable action profile, and the minimal curb set. Thus, analogous arguments apply when we weaken convergence to the set of correlated equilibria, iterative admissible action profiles, rationalizable action profiles, or minimal curb sets, respectively. □

A.3 Proof of Proposition 4

We show that there exists an uncoupled learning heuristic that in every game in **CI** leads to a Nash equilibrium with a payoff profile that strictly Pareto dominates any other payoff profiles. We do this by slightly modifying a learning heuristic used in Hart and Mas-Colell (2006, Proof of Theorem 3). Consider the following learning heuristic: for any $t \geq 2$,

- If $(a_R^{t-2}), a_C^{t-2}) = (a_R^{t-1}, a_C^{t-1})$, a_i^{t-1} is a best response to a_{-i}^{t-1}, and the payoff obtained by player i in $t-2$ and $t-1$ is player i's maximal payoff of the stage-game, then player i plays $a_i^t = a_i^{t-1}$.
- Otherwise, player i randomizes uniformly in t over all actions.

This learning heuristics is uncoupled; player i does not condition on the opponent's payoffs. It is easy to see that if both players follow the heuristic in any game in **CI**, then they will reach a pure Nash equilibrium corresponding to the strict Pareto dominant payoff profile almost surely. The arguments are analogous to Hart and Mas-Colell (2006, Proof of Theorem 3). This proves the first part of Proposition 4.

Let **U** be a product class of games such that each stage-game in **U** has a pure Nash equilibrium and **CI** $\subseteq$ **U**. Suppose by contradiction that there exists a profile of uncoupled learning heuristics $\boldsymbol{\sigma} = (\sigma_R, \sigma_C)$ that converges to Nash equilibrium in all games in **U** and is a Nash equilibrium learning heuristic of the learning game based on **U**. Consider the following stage-games:

$$\mathbf{u} = \begin{array}{c} \\ a \\ b \end{array} \begin{array}{c} \begin{array}{cc} a & b \end{array} \\ \begin{pmatrix} 8,9 & 2,3 \\ 0,1 & 4,5 \end{pmatrix} \end{array} \qquad \tilde{\mathbf{u}} = \begin{array}{c} \\ a \\ b \end{array} \begin{array}{c} \begin{array}{cc} a & b \end{array} \\ \begin{pmatrix} 8,9 & 6,3 \\ 0,7 & 4,5 \end{pmatrix} \end{array} \qquad \hat{\mathbf{u}} = \begin{array}{c} \\ a \\ b \end{array} \begin{array}{c} \begin{array}{cc} a & b \end{array} \\ \begin{pmatrix} 8,9 & 10,11 \\ 0,1 & 4,5 \end{pmatrix} \end{array}$$

We have $\mathbf{u}, \tilde{\mathbf{u}}, \hat{\mathbf{u}} \in$ **CI**.

Now consider

$$(u_R, \tilde{u}_C) = \begin{array}{c} \\ a \\ b \end{array} \begin{array}{c} \begin{array}{cc} a & b \end{array} \\ \begin{pmatrix} 8,9 & 2,3 \\ 0,7 & 4,5 \end{pmatrix} \end{array} \qquad (u_R, \hat{u}_C) = \begin{array}{c} \\ a \\ b \end{array} \begin{array}{c} \begin{array}{cc} a & b \end{array} \\ \begin{pmatrix} 8,9 & 2,11 \\ 0,1 & 4,5 \end{pmatrix} \end{array}.$$

Both games have a unique and pure Nash equilibrium. Since **U** is a product class and **CI** $\subseteq$ **U**, we have that both $(u_R, \tilde{u}_C), (u_R, \hat{u}_C) \in$ **U**.

Since (a,a) is the unique Nash equilibrium of $(u_R, \tilde{u}_C)$ and (b,b) is the unique Nash equilibrium of $(u_R, \hat{u}_C)$, the profile of learning heuristics $\boldsymbol{\sigma} = (\sigma_R, \sigma_C)$ must converge to (a,a) in $(u_R, \tilde{u}_C)$ and (b,b) in $(u_R, \hat{u}_C)$.

All of the above games are generic. Let $\tilde{\mathbf{U}}$ be an open neighborhood of $(u_R, \tilde{u}_C)$ and $\hat{\mathbf{U}}$ be an open neighborhood of $(u_R, \hat{u}_C)$ such that both are product classes and $\tilde{\mathbf{U}}, \hat{\mathbf{U}} \subseteq \mathbf{U}$ and retain the same pure best response structure, respectively.

Let $\tilde{\sigma}_C$ be an alternative learning heuristic for Colin that in games in $\tilde{\mathbf{U}} \cup \hat{\mathbf{U}}$ behaves identically to σ in $\tilde{\mathbf{U}}$. For all other games, $\tilde{\sigma}$ behaves identically to σ. That is, with $\tilde{\sigma}_C$ Colin pretends to be in $\tilde{\mathbf{U}}$ whenever he is $\hat{\mathbf{U}}$. Note that $(\sigma_R, \tilde{\sigma}_C)$ converges to (a,a) in stages-games in $\hat{\mathbf{U}}$, which is not the Nash equilibrium of these stage-games. Note further that Colin earns a strictly larger payoff in (a,a) than in the unique Nash equilibrium of these games. Since any nonempty open neighborhoods must have strict positive Lebesgue measure, Colin's average long run payoff from $\tilde{\sigma}_C$ in the learning game when Rowena follows σ_R is strictly larger than from σ_C. Thus, $\boldsymbol{\sigma}$ is not a Nash equilibrium of the learning game, a contradiction.

Note that we can extend the arguments above to games with larger actions sets by simply adding strictly dominated rows and columns.

Note further that we can let $\mathbf{U}$ be the smallest product class of games that contains **CI**, i.e.,

$$\mathbf{U} := \left\{ \mathbf{u} \in [0,1]^{2\cdot|A^2|} : \begin{array}{l} \mathbf{u} = (\tilde{u}_R, \hat{u}_C) \text{ s.t. there exist } \tilde{\boldsymbol{u}} = (\tilde{u}_R, \tilde{u}_C) \\ \text{and } \hat{\boldsymbol{u}} = (\hat{u}_R, \hat{u}_C) \text{ with } \tilde{\boldsymbol{u}}, \hat{\boldsymbol{u}} \in \mathbf{CI} \end{array} \right\}.$$

Finally note that both games, $(u_R, \tilde{u}_C)$ and $(u_R, \hat{u}_C)$, posses a unique, pure, and strict Nash equilibrium. Thus, it must also be the correlated equilibrium, the iterative admissible action profile, the rationalizable action profile, and the minimal curb set. Hence, analogous arguments apply when we weaken convergence to the set of correlated equilibria, iterative admissible action profiles, rationalizable action profiles, or minimal curb sets, respectively. This completes the proof of Proposition 4. □

A.4 Proof of Proposition 8

Define game $\mathbf{u}^3$ by

		Colin: a	Colin: b
Rowena	b	$1-2\varepsilon, \frac{1}{2} + \frac{\frac{1}{2}-\varepsilon}{2}$	$0,1$
	c	$1, \frac{1}{2} - \frac{\frac{1}{2}-\varepsilon}{2}$	$2\varepsilon, 0$

and $\mathbf{u}^4$ by

		Colin a	Colin b
Rowena	b	$1, \frac{1}{2} + \frac{\frac{1}{2}-\varepsilon}{2}$	$2\varepsilon, 1$
	c	$1-2\varepsilon, \frac{1}{2} - \frac{\frac{1}{2}-\varepsilon}{2}$	$0, 0$

Clearly there is $\hat{\varepsilon} > 0$ such that for all $\varepsilon \in [0, \hat{\varepsilon}]$, games $\mathbf{u}^3$ and $\mathbf{u}^4$ are generalizations of games $\mathbf{u}^1$ and $\mathbf{u}^2$ in Section 3, respectively, in the sense that they have identical pure best response structures, respectively. Colin's payoffs in $\mathbf{u}^3$ are identical to his payoffs in $\mathbf{u}^4$. Rowena's payoffs in $\mathbf{u}^3$ correspond to her payoffs in $\mathbf{u}^4$ except with rows being interchanged.

Note that (c, a) is the unique ε-Nash equilibrium of $\mathbf{u}^3$ and (b, b) is the unique ε-Nash equilibrium of $\mathbf{u}^4$. For Rowena, pretending to play $\mathbf{u}^3$ instead of $\mathbf{u}^4$ when both players follow a learning heuristic that converges $(1-\varepsilon)$ of the time to stage-game ε-Nash equilibrium is profitable if

$$(1-\varepsilon)\cdot(1-2\varepsilon) + \varepsilon \cdot 0 > \varepsilon \cdot 1 + (1-\varepsilon)\cdot 2\varepsilon$$

The first term of the l.h.s., $(1-\varepsilon)\cdot(1-2\varepsilon)$, is the payoff from stage-game ε-Nash equilibrium of $\mathbf{u}^3$ in $\mathbf{u}^4$, which happens at least $(1-\varepsilon)$ of the time. The second term of the l.h.s., $\varepsilon \cdot 0$, is the worst payoff that could be obtained in the remaining ε of the times when ε-Nash equilibrium is not played. The second term of the r.h.s. is the stage-game ε-Nash equilibrium payoff in $\mathbf{u}^4$, which happens at least $(1-\varepsilon)$ of the time. The first term of the r.h.s. is the highest payoff from non-Nash equilibrium of $\mathbf{u}^4$ which happens at most ε of the time. This is the best payoff in $\mathbf{u}^4$ to Rowena and hence the worst case for Rowena when she pretends to play $\mathbf{u}^3$ instead.

We solve above inequality together with $\varepsilon \in [0, 1)$ to find $\varepsilon \in \left[0, \frac{3}{4} - \frac{1}{4}\sqrt{5}\right)$. Set $\bar{\varepsilon} := \frac{3}{4} - \frac{1}{4}\sqrt{6}$. Clearly, $\bar{\varepsilon} > 0$. Thus, for all $\varepsilon \in [0, \bar{\varepsilon}]$ it is profitable for Rowena to pretend to play $\mathbf{u}^3$ instead of $\mathbf{u}^4$ when using a learning heuristic that converges at least $1-\varepsilon$ of the time to stage-game ε-Nash equilibrium and Colin uses an uncoupled learning heuristics converging at least $1-\varepsilon$ of the time to stage-game ε-Nash equilibrium.

Finally, since both $\mathbf{u}^3$ and $\mathbf{u}^4$ are generic, we can consider open neighborhoods $\mathbf{U}^3$ and $\mathbf{U}^4$ of $\mathbf{u}^3$ and $\mathbf{u}^4$, respectively, where this occurs. □

References

[1] Ashlagi, I., Monderer, D., and M. Tennenholtz (2006). Robust learning equilibrium, in: Dechter, R. and Richardson, T. (Eds.), Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence (UAI 2006).

[2] Aumann, R. and S. Sorin (1989). Cooperation and bounded recall, Games and Economic Behavior 1, 5–39.

[3] Babichenko, Y. (2010). Uncoupled automata and pure Nash equilibria, International Journal of Game Theory 39, 483–502.

[4] Başar, T. and G.J. Olsder (1999). Dynamic non-cooperative game theory, 2nd edition, SIAM.

[5] Basu, K. and J. Weibull (1991). Strategy subsets closed under rational behavior, Economics Letters 36, 141–146.

[6] Bernheim, B.D. (1984). Rationalizable strategic behavior, Econometrica 52, 1007-1028.

[7] Binmore, K. and L. Samuelson (1992). Evolutionary stability in repeated games played by finite automata, Journal of Economic Theory 57, 2780–305.

[8] Branfman, R.I. and M. Tennenholtz (2004). Efficient learning equilibrium, Artificial Intelligence 159, 27–47.

[9] Camerer, C. F., Ho, T.-H. and J.-K. Chong (2002). Sophisticated experience-weighted attraction learning and strategic teaching in repeated games, Journal of Economic Theory 104, 137–188.

[10] Chong, J.-K., Camerer, C. F., and T.-H. Ho (2006). A learning-based model of repeated games with incomplete information, Games and Economic Behavior 55, 340–371.

[11] Demichelis, S. (2013). Efficient coordination in repeated games: Behavioral maxims, mimeo.

[12] Duersch, P., Kolb, A., Oechssler, J. and B.C. Schipper (2010). Rage against the machines: How subjects learn to play against computers, Economic Theory 43, 407–430.

[13] Duersch, P., Oechssler, J., and B.C. Schipper (2012). Unbeatable imitation, Games and Economic Behavior 76, 88–96.

[14] Duersch, P., Oechssler, J., and B.C. Schipper (2014). When is Tit-for-tat unbeatable?, International Journal of Game Theory 43, 25–36.

[15] Ellison, G. (1997). Learning from personal experience: One rational guy and the justification of myopia, Games and Economic Behavior 19, 180–210.

[16] Ergin, H. and T. Sarver (2010). A unique costly contemplation model, Econometrica 78, 1285-1339

[17] Foster, D. and R. Vohra (1997). Calibrated learning and correlated equilibrium, Games and Economic Behavior 21, 40–55.

[18] Foster, D. and H.P. Young (2001). On the impossibility of predicting the behavior of rational agents, Proceedings of the National Academy of Sciences 98, 12848–12853.

[19] Foster, D. and H.P. Young (2003). Learning, hypothesis testing, and Nash equilibrium, Games and Economic Behavior 45, 73–96.

[20] Foster, D. and H.P. Young (2006). Regret testing: learning to play Nash equilibrium without knowing you have an opponent, Theoretical Economics 1, 341-367.

[21] Fudenberg, D. and D. Kreps (1993). Learning mixed equilibria, Games and Economic Behavior 5, 320 – 367.

[22] Fudenberg, D. and D. Levine (1999). Conditional universal consistency, Games and Economic Behavior 29, 104–130.

[23] Fudenberg, D. and D. Levine (1998a). The theory of learning in games, MIT Press.

[24] Fudenberg, D. and D. Levine (1998b). Learning and evolution: Where do we stand? Learning in games, European Economic Review 42, 631–639.

[25] Fudenberg, D. and D. Levine (1989). Reputation and equilibrium selection in games with a patient player, Econometrica 57, 759–778.

[26] Fudenberg, D. and E. Maskin (1990). Evolution and cooperation in noisy repeated games, American Economic Review Papers & Proceedings 80, 274–279.

[27] Germano, F. (2007). Stochastic evolution of rules for playing normal-form games, Theory and Decision 62, 311–333.

[28] Germano, F. and G. Lugosi (2007). Global Nash convergence of Foster and Young's regret testing, Games and Economic Behavior 60, 135–154.

[29] Hart, S. and A. Mas-Colell (2006). Stochastic uncoupled dynamics and Nash equilibrium, Games and Economic Behavior 57, 286–303.

[30] Hart, S. and A. Mas-Colell (2003). Uncoupled dynamics do not lead to Nash equilibrium, American Economic Review 93, 1830-1836.

[31] Hart, S. and A. Mas-Colell (2001). A General class of adaptive strategies, Journal of Economic Theory 98, 26–54.

[32] Hart, S. and A. Mas-Colell (2000). A simple adaptive procedure leading to correlated equilibrium, Econometrica 68, 1127–1150.

[33] Hyndman, K., Ozbay, E.Y., Schotter, A., and W.Z. Ehrblatt (2012). Convergence: An experimental study of teaching and learning in repeated games, Journal of the European Economic Association 10, 573–604.

[34] Israeli, E. (1999). Sowing doubt optimally in two-person repeated games, Games and Economic Behavior 28, 203–216.

[35] Jordan, J.S. (1991). Bayesian learning in normal form games, Games and Economic Behavior 3, 60–81.

[36] Kakade, S.M. and D.P. Foster (2008). Deterministic calibration and Nash equilibrium, Journal of Computer and System Sciences 74, 115–130.

[37] Kalai, E. and E. Lehrer (1993). Rational learning leads to Nash equilibrium, Econometrica 61, 1019–1045.

[38] Kim Y.K. (1994). Evolutionarily stable strategies in the repeated prisoner's dilemma, Mathematical Social Sciences 28, 167–197.

[39] Kordonis, I., Charalampidis, A., and G. Papavassilopoulos (2018). Pretending in dynamics games, alternative outcomes, and application to electricity markets, Dynamic Games and Applications 8, 844 – 873.

[40] Lemke, C.E. and J.T. Howson (1964). Equilibrium points of bimatrix games, Journal of the Society for Industrial and Applied Mathematics 12, 413–423.

[41] Lipman, B. (1991). How to decide how to decide how to ...: Modeling limited rationality, Econometrica 59, 1105–1125.

[42] Maschler, M., Solan, E., and S. Zamir (2013). Game theory, Cambridge University Press.

[43] Milgrom, P. and J. Roberts (1990). Rationalizability, learning, and equilibrium in games with strategic complementarities, Econometrica 58, 1255–1277.

[44] Monderer, D. and L.S. Shapley (1996). Potential games, Games and Economic Behavior 14, 124–143.

[45] Monderer, D. and M. Tennenholtz (2007). Learning equilibrium as a generalization of learning to optimize, Artificial Intelligence 171, 448–452.

[46] Mongin, P. and B. Walliser (1988). Infinite regressions in the optimizing theory of decision, in: Munier, B.E. (ed.), Risk, decision and rationality, D. Reidel Publishing Company, Dordrecht, 435–457.

[47] Moulin, H. (1979). Dominance solvable voting schemes, Econometrica 47, 1337–1351.

[48] Nachbar, J. (1997). Prediction, optimization, and learning in repeated games, Econometrica 65, 275–309.

[49] Nachbar, J. (2001). Bayesian learning in repeated games of incomplete information, Social Choice and Welfare 18, 303–326.

[50] Nachbar, J. (2005). Beliefs in repeated games, Econometrica 73, 459–480.

[51] Pearce, D. (1984). Rationalizable strategic behavior and the problem of perfection, Econometrica 52, 1029–1050.

[52] Sadzik, T. (2011). Coordination in learning to play Nash equilibrium, mimeo.

[53] Schipper, B.C. (2009). Imitators and optimizers in Cournot oligopoly, Journal of Economic Dynamics and Control 33, 1981–1990.

[54] Schipper, B.C. (2019). Dynamic exploitation of myopic best response, Dynamic Games and Applications 9, 1143–1167.

[55] Schmidhuber, J. (2006). Gödel machines: fully self-referential optimal universal problem solvers, in: B. Goertzel, B. and Pennachin, C. (Eds.), Artificial General Intelligence, Springer Verlag, 199–226.

[56] Shalev, J. (1994). Nonzero-sum two-person repeated games with incomplete information and known-own payoffs, Games and Economic Behavior 7, 246–259.

[57] Spohn, W. (1982). How to make sense of game theory, in: Stegmüller, W. Balzer, W., and Spohn, W. (Eds.), Philosophy of economics, Springer-Verlag, 239–270.

[58] Terracol, A. and J. Vaksmann (2009). Dumbing down rational players: Learning and teaching in an experimental game, Journal of Economic Behavior and Organization 70, 54–71.

[59] Young, H.P. (2009). Learning by trial and error, Games and Economic Behavior 65, 626–643.

[60] Young, H.P. (2004). Strategic learning and its limits, Oxford University Press.