Does Reasoning Enhance Learning? [*]

Nicolaas J. Vriend, *Universitat Pompeu Fabra*, Barcelona, Spain

October 1996

Abstract:

   Utilizing the well-known Ultimatum Game, this note presents the following phenomenon. If we start with simple stimulus-response agents, learning through naive reinforcement, and then grant them some introspective capabilities, we get outcomes that are not closer but farther away from the fully introspective game-theoretic approach. The cause of this is the following: there is an asymmetry in the information that agents can deduce from their experience, and this leads to a bias in their learning process.

J.E.L. classification codes: C7, D8

Keywords: Ultimatum Game, Game Theory, Reasoning, Reinforcement Learning

N.J. Vriend, Universitat Pompeu Fabra, Dept. Economics, Ramon Trias Fargas 25-27, 08005 Barcelona, Spain, <vriend@upf.es>, http://www.santafe.edu/~vriend/.

1. Introduction

One of the games most extensively studied in the literature during the last decade is the Ultimatum Game. The reason that this game is so intriguing seems to be that the game-theoretic analysis is straightforward and simple, while the overwhelming experimental evidence is equally straightforward but at odds with the game-theoretic analysis (see e.g., Güth et al. [1982], Güth & Tietz [1990], or Thaler [1988]).

Here is the basic form of the Ultimatum Game: there are two players, A and B, and a cake. Player A proposes how to split the cake between herself and player B. Upon receiving player A's proposal, player B has two options. First, to accept the proposal, which will then be executed. Second, to reject it, after which both get nothing. Many different variants of this basic setup have been considered in the literature, but the details are not essential for this note.

There are many Nash equilibria in this game. Every strategy for player A combined with any strategy for player B that accepts that offer but rejects all worse offers is one. But, considering the extensive form variant of the game, there is a unique subgame perfect equilibrium: player A offers the minimum portion, and player B accepts that.

Empirical evidence shows time and again that this is not what happens in most games played in the laboratory. Players A usually offer slightly less than half the cake to players B. And players B usually reject small offers. In this note we will concentrate on the behavior of players A, and assume that players B play the perfect equilibrium strategy. Concerning player A's behavior, there are two main explanations for the 'anomaly' offered in the literature. First, some argued that fairness and reciprocity considerations were the force driving players A to offer more than the game-theoretic approach would suggest to players B (see e.g., Forsythe et al. [1994]). A second explanation found in the literature is that players A are basically following an adaptive, best-reply seeking, approach to the behavior of players B. In a multi-period setup where players played the game repeatedly but against different players, some papers showed how it can happen that players A can unlearn to play the perfect equilibrium strategy before players B learn that they should play their perfect equilibrium strategy. And once players A do not play that strategy anymore, players B will never learn to play theirs (see Roth & Erev [1995] who apply a reinforcement learning approach, while Gale et al. [1995] use replicator dynamics).[1]

This note offers an additional possible explanation. If we start with simple stimulus-response agents, learning through naive reinforcement (see next section), and then grant them some introspective capabilities, we get outcomes that are not closer but farther away from the fully

---

[1] For a detailed comparative analysis of replicator dynamics and reinforcement learning, on which we focus in this note, see Börgers & Sarin [1995].

introspective game-theoretic approach. The cause of this is the following: there is an asymmetry in the information that agents can deduce from their experience, and this leads to a bias in their learning process. It is a simple point, but I have not seen it made in the literature. Section 2 will explain what reinforcement learning is, and the difference between naive reinforcement learning, and reinforcement learning plus reasoning. Section 3 shows how an information asymmetry leads to a distortion of the learning process. Section 4 discusses and relaxes the assumptions made in section 3, and section 5 will conclude.

2. Actual and Virtual Reinforcement Learning

Standard game theory is completely based on introspection. Given the agents' reasoning capabilities, actually playing games seems inessential.[2] If we imagine a rationality axis, the fully introspective players of standard game theory would be at one of the extremes of this axis. Close to the opposite end of the spectrum we would find boundedly rational agents without any reasoning capabilities. They act like conditioned animals, or like simple stimulus-response machines. These agents behave adaptively to their environment in that they experiment by trying actions, and actions that led in the past to better outcomes are more likely to be chosen again in the future. This type of behavior is nowadays known as reinforcement learning.

There is a family of stochastic dynamic models of such individual behavior in the scientific literature, for which different backgrounds can be distinguished. The idea was first developed in the psychological literature. See especially Hull [1943], and Bush & Mosteller [1955], on which Cross [1983] is based. Much later reinforcement learning was independently reinvented twice as a machine learning approach in computer science. See e.g., Barto at al. [1983], and Sutton [1992] for a survey. The other reinforcement learning approach in computer science is known as Classifier Systems. See Holland [1975] for early ideas on this, or Holland et al. [1986] for a much more elaborate treatment. In the economics literature reinforcement learning became better known only very recently through Roth & Erev [1995]. For expositional convenience, in this note we will follow the Roth & Erev approach to reinforcement learning.[3]

---

[2] Here is an unverified anecdote illustrating this. A famous economist who had just published a book about the game of Bridge, met a colleague who expressed his surprise about him being a player of this game. "I don't play Bridge", was his reaction, "I wrote a book about it".

[3] Reinforcement learning differs from some other important recent models of dynamics with experimentation in the economics literature (see e.g., Ellison [1993], Kandori et al. [1993], or Young [1993]). In those evolutionary dynamic models, adaptive behavior is basically a one-step error correction mechanism. The agents have a well-specified model of the game, they can reason what the optimal action would be, given the actions of the other players, completely independent from any payoff actually

In computer science the main reason for the interest in reinforcement learning is its success in performing difficult tasks. In psychology reinforcement learning is mainly judged on its success in explaining empirical evidence of human subjects in experiments. And the fact that this kind of model seems consistent with the robust properties of learning that have been observed in the large experimental psychology literature is also the main line of thought in the application of reinforcement learning models in the experimental economics literature (although Roth & Erev have some deeper ideas, on which more below).

Besides these reasons to take reinforcement learning serious, we take this general class of models as a starting point because it concerns the most basic form of learning. It is the simplest possible type of learning model, assuming the most limited cognitive capabilities, no reasoning or introspection at all, while still allowing for learning. Hence, besides the standard game-theoretic approach, this kind of model functions as a second bench-mark.

Consider the Ultimatum Game. Assume player A has no knowledge of game theory, the structure of the Ultimatum Game, or the behavior of players B. Player A is a boundedly rational agent who behaves adaptively to her environment. She simply tries actions, and is in the future more likely to repeat those actions that led to high payoffs in the past than those that led to low payoffs. One can imagine player A as playing with a multi-armed bandit, where different arms might give different payoffs, and player A does not know at the start which is the best arm to pull. This is the basic reinforcement learning approach (see Roth & Erev [1995]): At time t=1 each player has an initial propensity to play his kth pure strategy given by some real number $q_k(1)$. If a player plays his kth pure strategy at time t, and receives a payoff of z, then the propensity to play strategy k is updated by setting $q_k(t+1)=q_k(t)+z$, while for all other pure strategies j, $q_j(t+1)=q_j(t)$. The probability that the player plays his kth pure strategy at time t is $p_k(t) = qk(t)/\Sigma qj(t)$, where the sum is over all of the player's pure strategies j. We will call this *naive* reinforcement learning, or learning through actual reinforcement.

There are two obvious ways to extend the basic reinforcement learning mechanism. First, instead of learning only on the basis of his own actions and payoffs, an agent can learn from

---

[3](...continued)
experienced, and they play a best-response strategy against the frequency distribution of a given (sub-)population of other players. The evolutionary dynamics consist of a co-evolutionary adaptive process, players adapting to each others' adaptation to each other ..., plus experimentation in the form of trembling. The very first task for a reinforcement learning algorithm, on the other hand, is to learn what would be good actions. The agents do not have a well-specified model of their environment, and they do not know which action would be the best response. See Vriend [1994] for a more extensive discussion of these issues.

other agents' actions and payoffs. He can, then, reinforce his propensities to choose an action *as if* he had tried those actions, and experienced the payoffs himself. This is called learning through *vicarious* reinforcement (see Bandura [1977]).[4] We will not pursue that track here. Second, instead of learning on the basis of actions and payoffs actually experienced by himself or others, an agent can learn by reasoning about counterfactuals, i.e., about hypothetical actions and payoffs. He can, then, reinforce his propensities to choose an action *as if* he had actually tried those actions, and experienced the payoffs. We will call this learning through *virtual* reinforcement (see Holland [1990]).

Clearly, virtual reinforcement requires that one relaxes the restriction on player A's cognitive capabilities, and assumes that player A has some reasoning capabilities. In the Ultimatum Game the arms of the bandit can be ordered naturally from low offers to high offers. For example, when an offer x has been accepted by player B, player A can reason that player B would have accepted also all offers x'>x. Or when an offer x has been rejected, player A can reason that offers x"<x would have been rejected as well. Hence, besides the updating of the propensity to choose action x based on her actual experience, player A can also update the propensities to choose other actions based on the indicated simple reasoning process; *as if* she had actually tried those actions.

Adding the virtual updating will speed up player A's learning process. So far so good. The point, however, is that there is a second effect of the virtual updating, which is the following: There is an asymmetry in the information obtained by playing strategy x: if offer x is accepted, player A can reason what player B would have done with offers x'>x, but player A does not get information about what player B would have done with offers x"<x. Similarly, if offer x is rejected, player A can reason what player B would have done with offers x"<x, but not what player B would have done with offers x'>x.

3. Effect of the Information Asymmetry: the Basic Model

We will now analyze the effect of this information asymmetry on the reinforcement learning process. What we will do is start with some extremely simplifying assumptions to make our point. In section 4 the basic assumptions of table 1 will be discussed and relaxed.

---

[4] Notice that this is more sophisticated than imitation. It is only the propensity to choose an action that is updated on the basis of an other agent's action and consequence.

i)	The cake has size Π.	
ii)	The only possible offers x to player B are 0, 1, 2, ..., Π.	
iii)	Player B plays the perfect equilibrium strategy, and accepts every offer.[5]	
iv)	Player A tries every action equally often, say n times.	
v)	In case of acceptance the payoff to player A is simply Π-x. In case of rejection it would be 0.	
vi)	Reinforcement takes place as explained above: the payoff realized by playing offer x is simply added to the propensity to choose offer x.	
vii)	Only actual reinforcement learning takes place.	

Table 1 basic assumptions

Proposition 1: Under assumptions *i)* to *vii)*, the most reinforced offer will be x=0.

Proof: After player A has tried each possible offer n times, the propensity to choose a certain offer x will have increased with n times a payoff of $(\Pi-x)$. Formally, $r(x) = n \cdot (\Pi-x)$, which has a maximum at x=0. □

Next, we assume that player A reasons about counterfactuals as well. Hence we keep assumptions *i)* to *vi)*, but instead of *vii)*, we make the assumption given in table 2

vii')	Player A reasons that player B's behavior has a reservation value property. If offer x is accepted, offers x'>x would have been accepted too. Therefore, player A applies both actual and virtual reinforcement.

Table 2 assumption virtual reinforcement

Proposition 2: Under assumptions *i)* to *vi)* plus *vii')*, the most reinforced offer will be offer x satisfying $x>(\Pi-2)/2$ while $(x-1)\leq(\Pi-2)/2$.

Proof: After player A has tried each possible offer n times, the propensity to choose a certain offer x will have increased with n times a payoff of $(\Pi-x)$ through the actual reinforcement. In addition, the virtual reinforcement for an offer x will be: x times this actual reinforcement (we will give the intuition below). Hence, the total reinforcement for offer x is: $r(x) = n \cdot (x+1) \cdot (\Pi-x)$. Taking first differences gives: $r(x+1)-r(x) = n \cdot (\Pi-2x-2)$. Hence, $\Delta r(x)<0$ if $x>(\Pi-2)/2$. □

Notice that the only difference with the case of only actual reinforcement is the factor $(x+1)$. Since $(x+1)\geq1$, all propensities are updated more often now. That is, the virtual updating in addition to the actual updating makes the learning process faster. But there is also a second effect: the information asymmetry. Table 3 helps to explain this.

---

[5] To simplify the presentation we assume that a 0 offer is accepted as well.

offer x to player B	total reinforcement r(x)
0	$n \cdot (0+1) \cdot (\Pi-0)$
1	$n \cdot (1+1) \cdot (\Pi-1)$
2	$n \cdot (2+1) \cdot (\Pi-2)$
...	...
...	...
Π	$n \cdot (\Pi+1) \cdot (\Pi-\Pi)$

Table 3 actual plus virtual reinforcement

Consider the first line in table 3, i.e., the offer $x=0$. The propensity of this offer will be updated only when $x=0$ has actually been accepted. If other offers $x>0$ are accepted, player A will not know whether $x=0$ would have been acceptable too. Next, consider the offer $x=1$. This propensity will be reinforced actually when $x=1$ has been accepted, and in addition virtually when $x=0$ has been accepted. One can continue this up to the largest possible offer $x=\Pi$. If $x=\Pi$ has been accepted by player B, player A will update this propensity actually. But player A will also reason after every offer $x<\Pi$ that an offer $x=\Pi$ would have been acceptable as well. Hence, $x=\Pi$ will be updated $(\Pi+1)$ times as often as $x=0$. In other words, the first direct effect of the information asymmetry is that some offers are reinforced more often than other offers, although they are actually tried the same number of times. What does this mean for the final propensities to choose these offers? Formally, the total reinforcement for offer x is: $r(x) = n \cdot (\Pi-x) \cdot (x+1)$. Taking first differences gives: $r(x+1)-r(x) = n \cdot (\Pi-2x-2)$. Hence, $\Delta r(x)<0$ if $x>(\Pi-2)/2$. In other words, applying both actual and virtual reinforcement, the most reinforced offer x will be slightly below the 50-50 offer.

The example in figure 1 for $n=1$, and $\Pi=6$ illustrates this. The graph shows the difference between the two forms of learning. In one case only actual reinforcement takes place, while the other case allows for virtual reinforcement as well. In the first case the most reinforced offer is 0, while in the second case the most reinforced offers are 2 and 3.[6]

---

[6] Such offers are very common in laboratory experiments with a cake of size 6.
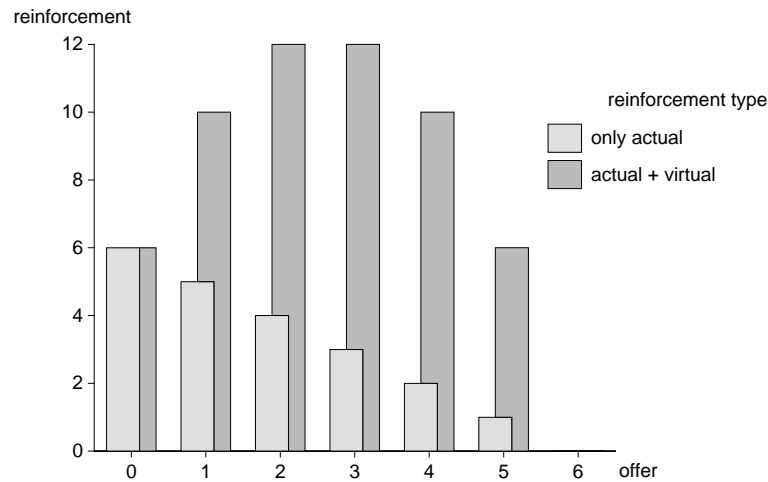
Figure 1 outcomes two types of reinforcement process

Before relaxing the assumptions we made in the next section, let us stress that the fact that offers slightly below half the cake size are most reinforced has nothing to do with fairness or reciprocity. Player A is assumed to be so boundedly rational that such concepts form no part of her cognitive capabilities. It has also nothing to do with best-responses to player B's behavior. In our exposition thus far player B accepts all offers made to him. The fact that offers slightly below half the cake size are most reinforced is due here entirely to the asymmetry in the information player A deduces from player B's acceptances.

4. Discussion of the Assumptions

As said above the simplifying assumptions made thus far were made in order to highlight the information asymmetry effect. We will now discuss these assumptions, and show how they can be relaxed without canceling this effect.

ad i) The are some papers in the literature in which the cake size Π is variable and not known with certainty (see e.g., Mitzkewitz & Nagel [1993]). We can allow for such a relaxation. It will not annul the effect of the asymmetry as such, but only change the numbers somewhat.

ad ii) Assuming that only integer offers up to Π can be made is without much loss of generality. As long as we assume there exists a minimum slice size, all that is involved is re-scaling that indivisible unit to 1.

ad iii) Assuming player B accepts all offers helps to distinguish the information asymmetry effect from the other explanations found in the literature concerning the empirical evidence. In general player B will not accept any offer; not even every offer with equal probability. But this

does not deny the asymmetry effect. And if, what seems most plausible, lower offers are rejected more often than higher offers, it only reinforces that effect (away from the perfect equilibrium). Adding the possibility of refusals, one should notice that there is a analogous asymmetry effect with refusals. If offer x to player B has been refused, player A can reason so would have been offers x"<x, but not necessarily offers x'>x. However, the standard assumption is that in case of refusal the payoff and hence the reinforcement is zero. Hence, this does not influence our result. We will come back to this point of reinforcement after refusals below.

ad iv) Player A will not try every strategy equally often, but will adapt towards the more reinforced ones. This relaxation will only reinforce the asymmetry effect. Making small offers to player B will get reinforced even less and less often.

ad v) Reinforcement to player A is not necessarily linear in the payoff in cake. It seems a harmless assumption. Changing it would change the numbers somewhat, but not cancel the asymmetry effect. If we want to take negative reinforcement in case of refusals into account as well, it only reinforces our argument, as it can be shown that the information asymmetry in case of rejections works in the same direction as the one discussed for the case of acceptance.

add vi) Suppose we do still do not update offers x"<x, but instead of making reinforcement additive, each rule's strength is averaged over the number of times it has been reinforced. Notice that implementing this relaxation is easy from a technical point of view. The question is, however, how realistic it is for human players. Additive reinforcement of the propensities is standard in the literature. Roth & Erev [1995] refer to evidence in the psychological literature (The Power Law of Practice; Blackburn [1936]) that supports this. Let us see what players are supposed to do when averaging instead of simply adding up. First, one should divide each rule's strength not only by the number of times it has actually been played, but as well by the number of times it has virtually been played and reinforced through the reasoning process. Second, all rules x"<x should be considered as not reinforced rather than as reinforced with factor 0. If the latter were done, each turn all propensities would be updated, and hence the averaging would make no difference. In order for the players to do this one needs to endow them with higher reasoning skills. Not only do they need to reason about counterfactuals. They also need to realize the existence of the asymmetry effect presented in this note, and they need to compensate correctly for it in their reinforcement.[7]

---

[7] Also, with averaging the asymmetry effect might come in as follows. Through the asymmetry effect, the higher offers would be updated relatively fast, their strengths 'walking' away from the initial values. Hence, you might get lock-in of the learning process.

ad vii') The result depends upon player A presuming player B's behavior has a reservation value property. One could relax this. As long as, after an acceptance of offer x, offers x'>x are presumed to be more likely accepted than offers x"<x, this does not deny the effect of the information asymmetry on the virtual updating as such, but only change the numbers somewhat. If player A may make inferences about offers x"<x as well, assigning probabilities to acceptance of those offers, and hence all offers are updated every time, then the asymmetry effect is avoided if and only if the agent forms correct estimates about the probability that such offers would have been accepted. However, the starting idea of an evolutive approach (instead of the fully introspective game-theoretic approach) was exactly that the agents have to learn such things during the game.[8]

5. Concluding Remarks

Before concluding with an observation and a conjecture, let me stress what this note is not about. First, this note is not meant to present virtual reinforcement as an improvement of the naive reinforcement learning model. The idea of virtual reinforcement as such is trivial. And there would be good reasons not to consider it an improvement of the basic reinforcement model. That model has been justified by the argument that besides the standard game-theoretic approach there is a need for low rationality game theory (see Erev & Roth [1995]). The basic idea is that agents do not understand how payoffs are determined, and perhaps do not even know what utility is, but they recognize it when they receive it. The question addressed, then, is for which classes of games can the experimental results be explained by such a low rationality model. This note is not to be meant critical to that approach at all. It is about the information asymmetry and its effect on the outcome of the reinforcement process once virtual updating would be considered. Second, whether players actually update virtually or not is an empirical question. But that is not our point. This note addresses the logical issue: whatever the extent in which players learn through virtual reinforcement, the effect of the information asymmetry is a bias away from the perfect equilibrium strategy. And if virtual reinforcement turns out to be relatively less utilized than actual reinforcement, this means that one could expect outcomes closer to the perfect equilibrium.

The exact contents of the propositions in this note are linked to the specific assumptions made. But the analysis of these propositions and assumptions leads to a paradoxical observation

---

[8] Moreover, relaxation here does not really matter as long as the reinforcement of offers x"<x is biased downwards as a result of the information asymmetry.

which is of more general relevance. Take traditional game theory and basic reinforcement learning, and consider them as two extremes on a rationality axis. The first based entirely on introspective processes, the latter abstracting completely from such processes.

Observation: Adding reasoning in the form of virtual reinforcement to the basic reinforcement learning model brings you closer to the fully introspective game-theoretic approach on the rationality axis. But it lead to outcomes that are farther away from the game-theoretic prediction.

In other words, we have shown that reasoning may *not* enhance learning. This result leads to the conjecture that the asymmetry in the information agents can deduce from their experience is of more general importance when boundedly rational agents are supposed to be endowed with some reasoning capabilities. The Ultimatum Game might, then, turn out to be merely a first example.

References
Bandura, A. (1977). *Social Learning Theory*. Englewood Cliffs, NJ: Prentice-Hall.
Barto, A.G., Sutton, R.S., & Anderson, C.W. (1983). Neuronlike Adaptive Elements That Can Solve Difficult Learning Control Problems. *IEEE Transactions on Systems, Man, and Cybernetics*, *13, No. 5*, 834-846.
Blackburn, J.M. (1936). *Acquisition of Skill: An Analysis of Learning Curves* (IHRB Report No. 73).
Börgers, T., & Sarin, R. (1995). *Learning through Reinforcement and Replicator Dynamics* (mimeo).
Bush, R.R., & Mosteller, F. (1955). *Stochastic Models for Learning*. New York: Wiley.
Cross, J.G. (1983). *A Theory of Adaptive Economic Behavior*. Cambridge: Cambridge University Press.
Ellison, G. (1993). Learning, Local Interaction, and Coordination. *Econometrica*, *61*, 1047-1071.
Erev, I., & Roth, A.E. (1995). *On the Need for Low Rationality, Cognitive Game Theory: Reinforcement Learning in Experimental Games with Unique, Mixed Strategy Equilibria* (mimeo).
Forsythe, R.J., Horowitz, J.L., Savin, N.E., & Sefton, M. (1994). Fairness in Simple Bargaining Experiments. *Games and Economic Behavior*, *6*, 347-369.
Gale, J., Binmore, K., & Samuelson, L. (1995). Learning to Be Imperfect: The Ultimatum Game. *Games and Economic Behavior*, *8*, 56-90.
Güth, W., Schmittberger, R., & Schwartz, B. (1982). An Experimental Analysis of Ultimatum Bargaining. *Journal of Economic Behavior and Organization*, *3*, 367-388.
Güth, W., & Tietz, R. (1990). Ultimatum Bargaining Behavior: A Survey and Comparison of Experimental Results. *Journal of Economic Psychology*, *11*, 417-440.
Holland, J.H. (1975). *Adaptation in Natural and Artificial Systems*. Ann Arbor: University of Michigan Press.
Holland, J.H. (1990). Concerning the Emergence of Tag-mediated Lookahead in Classifier Systems. *Physica D*, *42*, 188-201.
Holland, J.H., Holyoak, K.J., Nisbett, R.E., & Thagard, P.R. (1986). *Induction: Processes of Inference, Learning, and Discovery*. Cambridge, MA: MIT Press.
Hull, C.L. (1943). *Principles of Behavior*. New York: Appleton-Century-Crofts.
Kandori, M., Mailath, G.J., & Rob, R. (1993). Learning, Mutation, and Long Run Equilibria in Games. *Econometrica*, *61*, 29-56.
Mitzkewitz, M., & Nagel, R. (1993). Experimental Results on Ultimatum Games with Incomplete Information. *International Journal of Game Theory*, *22*, 171-198.

Roth, A.E., & Erev, I. (1995). Learning in Extensive-Form Games: Experimental Data and Simple Dynamic Models in the Intermediate Term. *Games and Economic Behavior*, *8*, 164-212.

Sutton, R.S. (1992). Introduction: The Challenge of Reinforcement Learning. *Machine Learning*, *8, Nos. 3/4*, 225-227.

Thaler, R.H. (1988). The Ultimatum Game. *Journal of Economic Perspectives*, *2*, 195-206.

Vriend, N.J. (1994). Artificial Intelligence and Economic Theory. In E. Hillebrand & J. Stender (Eds.), *Many-Agent Simulation and Artificial Life* (pp. 31-47). Amsterdam: IOS.

Young, H.P. (1993). The Evolution of Conventions. *Econometrica*, *61*, 57-84.