ELSEVIER
Journal of Economic Behavior & Organization
Vol. xxx (2006) xxx–xxx
JOURNAL OF
Economic Behavior
& Organization
www.elsevier.com/locate/econbase

Evolution of cooperation in a one-shot Prisoner's Dilemma based on recognition of trustworthy and untrustworthy agents

Marco A. Janssen

School of Human Evolution and Social Change, Department of Computer Science and Engineering, Arizona State University, Box 872402, Tempe, AZ 85287-2402, USA

Received 6 July 2004; accepted 16 February 2006

Abstract

This article explores the conditions under which agents will cooperate in one-shot two-player Prisoner's Dilemma games if they are able to withdraw from playing the game and can learn to recognize the trustworthiness of their opponents. When the agents display a number of symbols and they learn which symbols are important to estimate the trustworthiness of others, agents will evolve who cooperate in games in line with experimental observations. These results are robust to significant levels of mutations and errors made by the players.

JEL classification: B52; C70

Keywords: Estimating trustworthiness; Cooperation; One-shot Prisoner's Dilemma

1. Introduction

Several theories have been proposed to explain the evolution of human cooperation. The theory of kin selection focuses on cooperation among individuals who are closely related genetically (Hamilton, 1964), whereas theories of direct reciprocity focus on the selfish incentives for cooperation in repeated interactions (Trivers, 1971; Axelrod, 1984). The theories of indirect reciprocity and costly signaling show how cooperation in larger groups can emerge when the cooperators can build a reputation (Alexander, 1987; Nowak and Sigmund, 1998; Lotem et al.,

E-mail address: Marco.Janssen@asu.edu.

0167-2681/$ – see front matter © 2006 Published by Elsevier B.V.
doi:10.1016/j.jebo.2006.02.004

1999; Wedekind and Milinski, 2000; Leimar and Hammerstein, 2001; Zahavi, 1977; Gintis et al., 2001).

Cooperation in repeated settings is well explained, but as Field (2001) points out, cooperation in one-shot Prisoner Dilemma games is an important finding of experimental research that needs to be understood to improve insight to the evolution of human behavior. Altruistic punishment is proposed as an explanation of why cooperation is found frequent among genetically unrelated people, in non-repeated interactions, when gains from reputation are small or absent (Fehr and Gächter, 2002). In this article I propose an alternative explanation, namely the ability to recognize untrustworthy opponents. Trusting behavior can explain cooperation in experimental one-shot games (Ostrom and Walker, 2003). Recognition of trustworthiness of others has been found to be an important factor in experimental studies (Frank et al., 1993; Mealey et al., 1996; Oda, 1997). Humans especially have an ability to recognize defectors (see also Cosmides, 1989).

Through the use of simulation experiments with artificial agents, I will show that evolution of cooperation in non-repeated interactions between unrelated agents might be possible when the agents have the ability to learn to recognize the trustworthiness of other agents. Also, the agents have different tendencies to cooperate and will learn to recognize these agent types. An analytical model of one-shot games in relation to honest and dishonest players has been developed by Frank (1987). He analyzed equilibria for different assumptions of the information from signals and costs of adapting the signals. The difference between this study and Frank's study is the ability not to play a game and the explicit learning process of the agents. Costs of detecting the types of players, constant in Frank, is a function of learning abilities as I will discuss later in this article. The model also relates to the indirect evolutionary approach (Güth and Kliemt, 1998; Berninghaus et al., 2003) in which subjective and objective payoffs explicitly different play a role. A subjective pay-off function and an objective pay-off function apply simultaneously. The former is driving choice, the latter selection. Güth and colleagues use evolutionary game theory and do not distinguish interactions between individuals but apply fractions of agent types. They show that a mixed equilibrium of trustworthy and untrustworthy agents can evolve.

The model builds on the literature of partner selection and the use of symbols in Prisoner's Dilemma games to recognize types of agents and show that cooperation occurs when agents cooperate only with agents with the same symbols (Hales, 2001; Lindgren and Nordahl, 1994; Riolo, 1997; Riolo et al., 2001). Macy and Skvoretz (1998) use symbols related to the types of players' behavior.

Besides providing the decision to cooperate or defect in a Prisoner's Dilemma, various scholars have included the option of not playing the game, in experimental tests as well as modeling exercises. Orbell et al. (1984), Orbell and Dawes (1991) and Hauk and Nagel (2001) argue that players with the intention to defect have a higher level of not playing the game compared with players who have an intention to cooperate. In model analyses of Schluessler (1989) and Vanberg and Congelton (1992), agents with successful strategies are those who exit when the opponent defected in the previous game. In fact, these strategies do not tolerate errors. Stanley et al. (1994) and Ashlock et al. (1996) allowed players to chose partners, and their chosen partners to refuse offers, based on the known history of interactions with other players.

In the next section I will present my model that combines ideas from partner selection and the use of tags in order to study whether agents learn to recognize trustworthiness of strangers in playing one-shot Prisoner's Dilemma games. In Section 3, I will discuss the experimental set-up, the basic results are presented in Section 4, and a sensitivity analysis in Section 5. The conclusions are presented in Section 6.

Please cite this article as: Marco A. Janssen, Evolution of cooperation in a one-shot Prisoner's Dilemma based on recognition of trustworthy and untrustworthy agents, Journal of Economic Behavior & Organization (2006), doi:10.1016/j.jebo.2006.02.004

2. The model

The model consists of a population of n players who randomly play one-shot two-person Prisoner's Dilemma games. I define the game they play, which strategy the players use, what types of symbols the players display, and how they learn to recognize the trustworthiness of these symbols. Finally, I discuss how the composition of players evolves when the game is played for a number of generations.

2.1. The game

Each individual has three possible actions: cooperate (C), defect (D), or withdraw (W). If both players cooperate, they each get a payoff of R (reward for cooperation). If both players defect, they each get a payoff of P (punishment for defecting). If player A defects and B cooperates, A gets a payoff of T (temptation to defect), and B gets S (sucker's payoff). If at least one of the players withdraws from the game, both players get a payoff of E (exit payoff). The resulting payoffs are given in Table 1.

In line with Tullock (1985) and Vanberg and Congelton (1992), I assume that the costs and benefits of exit are such that the expected payoffs from choosing not to play are higher than those resulting from mutual defection, but lower than those expected from mutual cooperation. The Prisoner's Dilemma is defined when $T > R > E > P > S$ and $2R > T + S$. In this situation the best option for any one move is to withdraw from the game. If one expects that the other agent will cooperate, the best option is to defect. If one expects that the other agent will defect, the best option is to withdraw. Since the game is symmetrical, each player comes to the same conclusion, so they both withdraw and end up with payoffs that are much lower than if they both trust that the other will cooperate. The pay-off matrix for the game in this article is defined using $T = 2$, $R = 1$, $E = 0$, $P = -1$, and $S = -2$, which is in line with Tullock.

Experimental research has shown that the material payoff does not have to equal the utility payoff experienced by the players (see, for example, Ahn et al., 2001, 2003). Not every subject shows selfish behavior. In fact, the majority of non-selfish players seem to be conditionally cooperative and cooperate only when they know that the other will probably cooperate. I will use the notion of social-welfare preferences to formulate the utility of agents (e.g. Andreoni and Miller, 2002). With these social-welfare preferences, subjects always prefer more for themselves and the other person, but are more in favor of getting payoffs for themselves when they are behind than when they are ahead. The strength of such preferences is increasing in the magnitudes of parameters α and β. The utility can then be formulated as:

$$u_i = \pi_i - \alpha_i \max(\pi_i - \pi_j, 0) + \beta_i \max(\pi_j - \pi_i, 0) \qquad (1)$$

Table 1
Pay-off table of the Prisoner's Dilemma with the option to withdraw from the game

	Player B		
	Cooperate	Defect	Withdraw
Player A			
Cooperate	R, R	S, T	E, E
Defect	T, S	P, P	E, E
Withdraw	E, E	E, E	E, E

Please cite this article as: Marco A. Janssen, Evolution of cooperation in a one-shot Prisoner's Dilemma based on recognition of trustworthy and untrustworthy agents, Journal of Economic Behavior & Organization (2006), doi:10.1016/j.jebo.2006.02.004

Table 2
Utility pay-off table of the Prisoner's Dilemma with the option to withdraw from the game

	Player B		
	Cooperate	Defect	Withdraw
Player A			
Cooperate	R, R	$S+\beta_{\mathrm{A}}(T-S), T-\alpha_{\mathrm{B}}(T-S)$	E, E
Defect	$T-\alpha_{\mathrm{A}}(T-S), S+\beta_{\mathrm{B}}(T-S)$	P, P	E, E
Withdraw	E, E	E, E	E, E

where u_i is utility of agent i, and π_i the monetary income of agent i. We define $\beta_i \leq \alpha_i$ and $0 \leq \beta \leq 1$. The α value can be regarded as the strength of an individual's aversion to exploiting others, and β can be regarded as an individual's degree of altruistic tendency. These α and β values determine the strategies of the agents to cooperate or not. Furthermore, the simulated evolutionary process may affect the α and β values of the agents. Both aspects will be explained below in more detail. Although there are other competing utility models of other regarding preferences, Charness and Rabin (2002) find that social-welfare preferences explain the data best for a comprehensive set of experimental data. The experimental estimates for α and β vary around 0.4 and 0.1.

To include the heterogeneity of motivations, I formulate the utility function in Table 2, where the material payoffs can be adjusted by individual motivations.

An agent has three elements: (1) the set of symbols that it displays, (2) the strategy that it uses to decide whether to trust or not another agent, and (3) the strategy it uses in Prisoner's Dilemma games.

The symbols are represented in the following way. Each agent has s symbols that can have value 0 or 1, where 0 means no display of the symbol and 1 means display of the symbol. The other two elements require more discussion.

2.2. *Trust*

The rule an agent uses to decide to trust the other agent, and thus be willing to play a Prisoner's Dilemma game, is represented as a single-layer neural network (Janssen and Ostrom, 2006). Such a basic neural network model is a simple but effective model to represent human learning processes of pattern recognition (Mehrota et al., 1997). A neural network has s inputs, which are the values 0 and 1 of the other agent's symbols. A weighted sum M of these inputs is calculated using the following equation:

$$M = w_0 + \sum_{i=1}^{s} w_i x_i, \tag{2}$$

where w_0 is the bias, w_i the weight of the ith input, and x_i the ith input. Initially, all weights are zero, but during the simulation the network is trained, when new information is derived, by updating the weights as described below in Eq. (4).

Neural networks use a so-called threshold function to translate the inputs into one output. The standard threshold function used for neural networks is a sigmoid function, and it determines trust

Please cite this article as: Marco A. Janssen, Evolution of cooperation in a one-shot Prisoner's Dilemma based on recognition of trustworthy and untrustworthy agents, Journal of Economic Behavior & Organization (2006), doi:10.1016/j.jebo.2006.02.004

defined as the probability Pr[Tr] that the agent will cooperate with its prospective partner:

$$\Pr[\mathrm{Tr}] = \frac{1}{1+e^{-M}}. \tag{3}$$

The higher the value of M, the higher the probability will be. The probability of not trusting the other agent is $1-\Pr[\mathrm{Tr}]$. Since the initial weights are assumed to be zero, the initial value of Pr[Tr] is 0.5.

If a game is played, each agent receives feedback, F, on the experience. This feedback is simply whether the partner cooperated or not. If the partner cooperated ($F=1$), the agent adjusts the weights associated with the other agent's symbols upward, so that it will be more likely to trust that agent, and others displaying similar symbols, in the future. On the other hand, if the partner defected ($F=0$), the agent will adjust the same weights downward, so that it will be less likely to trust that agent and others with similar symbols. The equation to adjust the weights is as follows:

$$\Delta w_i = \lambda(F - \Pr[\mathrm{Tr}])x_i, \tag{4}$$

where Δw_i is the adjustment to the ith weight, λ the learning rate, F the feedback, $F-\Pr[\mathrm{Tr}]$ the difference between the agent's level of trust in the other agent and the observed trustworthiness of the other agent, and x_i the other agent's ith symbol. In effect, if the other agent displays the ith symbol, the corresponding weight is updated by an amount proportional to the difference between the observed trustworthiness of an agent and the trust placed in that agent. The weights of symbols associated with positive experiences increase, while the weights of those associated with negative experiences decrease, reducing discrepancies between the amount of trust placed in an agent and that agent's trustworthiness.

The initial values of α are drawn from a uniform distribution between 0 and 1, and for β between 0 and 1, by which only initial conditions are accepted where $\beta \leq \alpha$. The initial values of the symbol x_i, 0 or 1 for each, are chosen randomly, and all initial weights w_i are set to 0.

2.3. Strategies

When neither agent withdraws from playing a game, they have to decide to cooperate or to defect. The agents are assumed conditionally to cooperate. In Vanberg and Congelton (1992), conditional cooperation was deterministic; as soon as the other player defects, the agent will withdraw from playing the game. Since in my case the players play one-shot games, such a strategy will not work. I assume that the players will estimate the expected utility for cooperation, $E[U(C)]$, or defection, $E[U(D)]$. The expected utility is determined by assuming that the level of expected trust of an agent in its opponent, defined in (3), represents the probability that the opponent will cooperate:

$$E[U(C)] = \Pr[\mathrm{Tr}]R + (1-\Pr[\mathrm{Tr}])(S + \beta_i(T-S)) \tag{5}$$

or

$$E[U(D)] = \Pr[\mathrm{Tr}](T - \alpha_i(T-S)) + (1-\Pr[\mathrm{Tr}])P. \tag{6}$$

Given the two estimates of expected utility, the player is confronted with a discrete choice problem, which I address with a logit function. The probability to cooperate, Pr[C], depends on the expected utilities and the parameter γ, which represents how sensitive the player is to differences in the estimates. The higher the value of γ, the more sensitive the probability to cooperate is to differences between the estimated utilities:

$$\Pr[C] = \frac{e^{\gamma E[U(C)]}}{e^{\gamma E[U(C)]} + e^{\gamma E[U(D)]}}. \quad (7)$$

Logit models are used by other scholars to explain observed behavior in one-shot games such as the logit equilibrium approach by Anderson et al., in press (forthcoming). Although the functional relation is similar, their approach differs from mine since they assume an equilibrium and perfect information of the actions and motivations of other players. Moreover, in my games agents do not play anonymously but can observe symbols of others in order to estimate the behavior of the opponent.

2.4. Generations

In one generation a certain number (g) of games are played. For each of these, two agents are chosen at random. These agents then decide if they trust each other by the process described in Section 2.2. If they do trust each other, they play the Prisoner's Dilemma once (see Section 2.3) and the resulting payoffs are added to their respective total scores.

The average material payoff an agent has received from all its interactions (games played or games exited) with other agents is used to determine the offspring in the next generation. Each generation 10% of the population is replaced by new agents who are selected from the population (including those who previously left the system) using a tournament selection algorithm. Two contestants are picked at random from the population, and their average payoffs are compared. The one with the higher average payoff becomes a new agent. If both contestants have identical scores, the winner is picked at random.

The genotype of the new agent (α, β, x_i, w_i) is copied from the parent and then may experience mutation. The mutation rates for strategies, symbols, and weights determine the probability or degree that each individual component will be mutated. In the case of the symbol, a mutation consists of flipping the component to the opposite state (0 (off) to 1 (on) or vice versa). In the case of the motivations and weights, there is always a mutation draw from a Gaussian distribution with means equal to the parameter values after crossover. The level of mutation is set by the standard deviations of the Gaussian distributions (Table 3).

Note that the model includes two types of adaptation of the weights. Within a generation agents are able to update the weights in order to learn to recognize the agent types. Between generations agents derive weights from the previous generation, although those weights are somewhat altered by mutations.

Table 3
List of parameters and their default values

Parameter	Value
Number of agents (n)	100
Number of symbols (s)	20
Learning rate (λ)	1.0
Steepness (γ)	2
Number of games per generation (g)	500
Mutation rate symbols	0.05
Standard deviation mutation w, α, β	0.1

3. Experimental design

Conventional rational choice theory predicts that selfish players will evolve toward avoiding the game. I will show that the inclusion of recognition leads to a population of agents that are not selfish and reach high levels of cooperation.

Parameter values for the default case are presented in Table 3. I will use this default case as a reference for testing the sensitivity for a number of assumptions.

Each set of parameter values is used to simulate 50 runs, where each run consists of 1500 generations. Taking into account an initialization period, I report statistics for the last 1000 generations. I report the average and standard deviation of withdraw, cooperation, defection, payoffs, the combinations of mutual cooperation, mutual defection, and cooperation/defection. I will analyze the influence of the number of symbols, the number of players, the learning rate, the mutation rates, and the error probabilities.

4. Basic results

A typical experiment of the default case is depicted in Fig. 1. Initially, agents usually withdraw from playing a game, but they learn quickly who to trust in playing a game and how to play it. Fig. 1 shows that after about 200 generations cooperation reaches a high level, although periodic collapses of cooperation can occur. These periodic collapses are caused by invasions of defectors who express symbols that are recognized by others as symbols of trustworthiness. Such defectors can emerge due to mutations and spread rapidly in a population of agents who recognize them as being trustworthy. After a number of generations, the exploitive behavior of the defectors is noticed, and cooperative agents have learned to recognize them, leading to an increase in the level of cooperation within the total population.

Table 4 shows the basic statistics of the default case of 50 runs under the same conditions. The results show that a level of cooperation is established (around 80%), which is significantly different from the conventional prediction that all players will withdraw from playing the games. The average value of parameter α is 0.75, and there is an intention to be altruistic, with β around 0.4. The standard deviation of β is relatively large, 0.16. The reason for this is that $\alpha \geq \beta$, and that

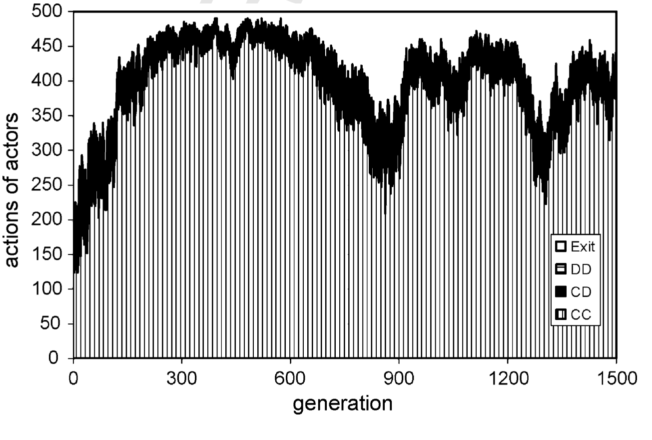

Fig. 1. Number of agents engaged in mutual cooperation (CC), mutual defection (DD), cooperation/defection (CD) or withdrawal from playing the game (Exit) for each generation.

Please cite this article as: Marco A. Janssen, Evolution of cooperation in a one-shot Prisoner's Dilemma based on recognition of trustworthy and untrustworthy agents, Journal of Economic Behavior & Organization (2006), doi:10.1016/j.jebo.2006.02.004

Table 4
Basic statistics of the default case with the mean and standard deviation (parentheses) for 50 runs

Indicator	Value
Payoff	0.81 (0.15)
# CC games	405.7 (74.9)
# CD games	25.0 (9.4)
# DD games	0.7 (0.6)
# Withdrawn	68.6 (72.5)
Parameter α	0.76 (0.10)
Parameter β	0.41 (0.18)

when $\alpha > 0.25$ the agent will value cooperation over defection. Therefore, the value of β is less influential than α and might therefore fluctuate more due to lower selection pressure.

The average values of α is higher than the empirical estimates in Charness and Rabin (2002), but in empirical studies the level of cooperation in one-shot experiments is 50% instead of our 80%. We will explore in the following section how relaxing assumptions lowers the cooperation levels, which are more in line with empirical observations.

In contrast to Frank (1987), my agents do not pay a price to get correct information on what type of player the opponent is. The model of Frank predicts a 100% level of cooperation in such a case, where 80% is reached in my simulations. The explanation is that a price is paid indirectly for knowing the information of the opponent. Namely, they have to learn by trial and error, which symbols are good predictors of trustworthiness. Since symbols related to the types of agents slowly evolve over time, agents need to remain exposed to defectors in order to update their weightings of the symbols. Furthermore, there is a cost of exploration since mutations have the probability of changing cooperators into defectors.

Recently, Skyrms (2002) and Miller et al. (2002) explored the consequences of "cheap talk", costless signaling, on cooperation in Dilemma situations. Skyrms looks at stag hunt games and bargaining games, while Miller et al. (2002) look at one-shot Prisoner Dilemma games. In both studies cooperation can evolve, but this situation is unstable and rapidly disappears again. Skyrms (2002) and Miller et al. (2002) assume that agents are selfish rational agents. I assume agents are also rational, but due to evolutionary pressures, the average social-welfare preferences increase. The fact that agents can decide not to play a game when they do not trust their opponent is instrumental in the evolution of other social-welfare preferences. When the exit option is not possible, no cooperation can evolve. This can be understood by the following reasoning. Suppose you are a trustworthy agent who has to make a decision of cooperation or defection in a game with a stranger. Suppose you do not trust your opponent; what do you do? If you cooperate, since you are trustworthy, you can expect the sucker payoff, but if you defect, the opponent will view you as non-trustworthy and adjust its weights for estimating trustworthiness of strangers. As a consequence, agents evolve to defect frequently, and there is not a strong selection pressure to weed out non-trustworthy agents. Hence the inclusion of the exit option and the evolution of social-welfare preferences are crucial assumptions leading to my results.

5. Sensitivity analysis

The number of symbols can affect the level of cooperation. Fig. 2 shows that a reduction of the number of symbols to below 15 significantly reduces the average payoff. The average

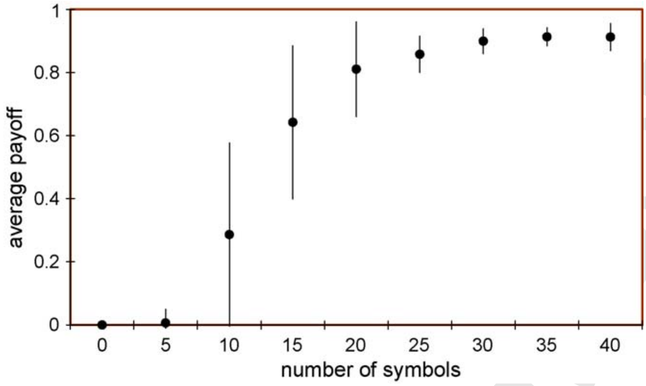

Fig. 2. The effect of the number of symbols on the average payoff. The dots refer to the average value of 50 runs; the lines represent the standard deviation of these 50 observations.

payoff reaches zero around five symbols, which can be interpreted to mean that there is not enough diversity in symbols to distinguish trustworthy and non-trustworthy agents. Increasing the number of symbols beyond 20, as in the default case, only slightly increases the average payoff. There is a decreasing marginal return to the number of symbols. Note that for seven symbols all agents can have a unique representation. In the simulation the agents are descendants affected by mutations. This requires more symbols to derive enough diversity in the simulated population.

The next question is whether or not the results are the consequence of agents learning to recognize trustworthiness in others. To test this, I varied the learning rate λ between 0 and 1. The resulting statistics, depicted in Fig. 3, show that when the learning rate is below 0.3 the average payoff drops significantly. When the learning rate is zero, the average payoff is zero, where almost no agents play the game. These results show that cooperation in my model is sensitive to the ability to learn to recognize whom to trust. Note that a high learning rate does not led to 100% cooperation. The reason is that agents need to be confronted with defectors in order to update the recognition ability of recognizing agent types with changed symbolic expressions.

In the previous experiments I draw the initial distribution of α and β between 0 and 1. We investigated the effect of starting with different groups of players by sampling a narrower margin of the parameter α. Fig. 4 shows that starting with agents biased toward selfish behavior results

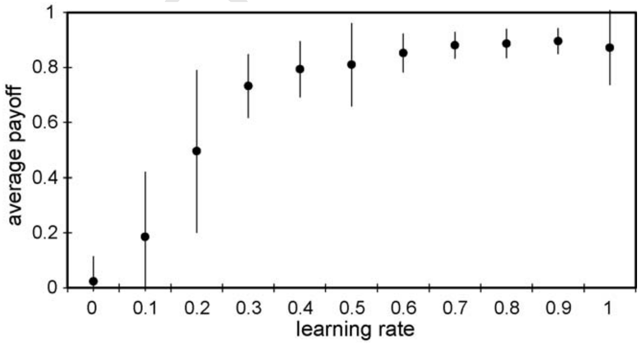

Fig. 3. The effect of the learning rate on the average payoff. The dot represents the average of 50 simulations; the lines represent the standard deviation.

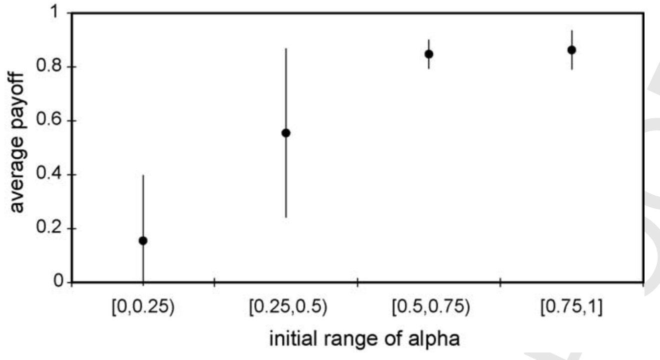

Fig. 4. The effect of the initial distribution of α on the average payoff after 1500 generations. The dot represents the average of 50 simulations; the lines represent the standard deviation.

in a modest average payoff level of 10%. Given the high variance of the outcomes, it shows that groups can evolve into more cooperative agents. If we let the model run 10,500 generations instead of 1500 generations, the average payoff eventually reaches 80% when the initial range of α is between 0 and 0.25. Only when α is sampled from values between 0.5 and 1 does a stable level of cooperation evolve within 1500 generations. I found that different initial samplings for β did not affect the results. As explained before, the value of α is dominant whether the agent has motivation to defect or not. In that result, the effect of α is dominant, at least in the short run, since simulations (not shown) illustrate that after many generations the cooperation levels converge to 80% independent of the starting values.

Thus far, I have assumed that agents make their decisions probablistically but do not make mistakes. Suppose, for example, the agent tosses a biased coin to decide whether or not to trust the agent and makes a mistake in reading the coin. How sensitive is the level of cooperation to mistakes? In Fig. 5 an increase is observed in probability p_{T}, the mistake of trusting an untrustworthy opponent or the mistake of distrusting a trustworthy opponent, leading to a lower level of payoff; this decrease is almost linear with the increase in the value of p_{T}.

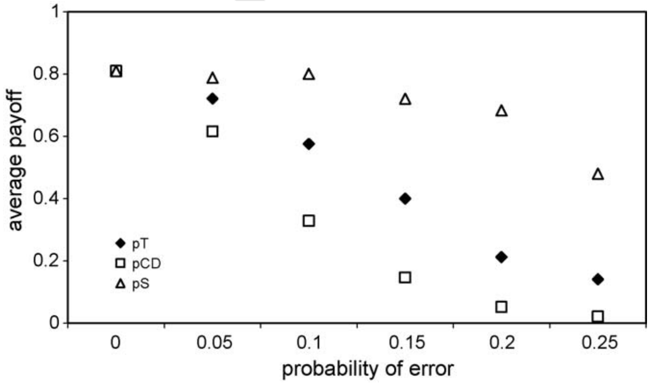

Fig. 5. The effect of the probabilities p_{T}, p_{CD} and p_{S} of making mistakes in trusting the opponent, deciding to cooperate or defect, and reading a symbol on the average payoff.

Please cite this article as: Marco A. Janssen, Evolution of cooperation in a one-shot Prisoner's Dilemma based on recognition of trustworthy and untrustworthy agents, Journal of Economic Behavior & Organization (2006), doi:10.1016/j.jebo.2006.02.004

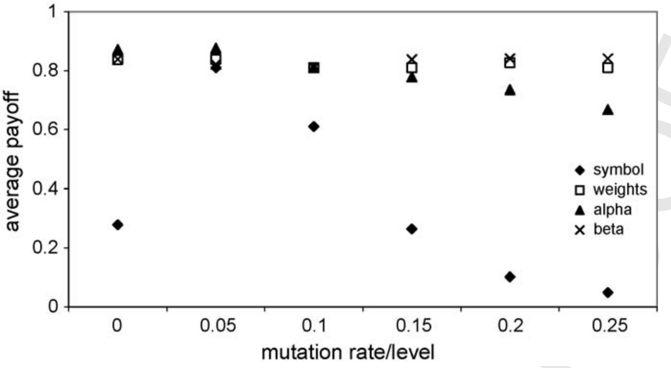

Fig. 6. The relationship between mutation rates for the symbols, the weights, and the social welfare preferences parameters, and the average payoff.

In a similar way the impact is assessed of making a mistake in deciding to cooperate. If an agent would normally cooperate, based on tossing the coin with probability as defined in Eq. (7), then the agent may make a mistake with probability p_{CD}. The results are relatively sensitive to these types of mistakes: a value of p_{CD} above 0.1 will reduce the average payoff to nearly zero. Making such mistakes is costly. Since most of the players are eager to cooperate, such mistakes are unintended defections, and this will affect the expected trust of opponents in an agent's symbols. This reduces the opportunities for the agent and its offspring to play games.

The third type of mistake is to misread the symbol, for example by overlooking a symbol of the opponent, which affects the estimated trustworthiness. The results shown in Fig. 5 illustrate the minor impact of this type of error. An explanation is that there is redundancy in the list of symbols that makes the estimation of trustworthiness robust for a certain degree of mistakes, but this does not hold for making errors in trusting other agents (p_{T} and p_{CD}) since such errors directly affect the level of cooperation.

The characteristics of the agents change over time due to mutations. This can be interpreted in two different ways. First, as a learning process, where agents copy the characteristics of successful other agents. Mutation rates for symbols and mutation levels for weights and social welfare preferences are in this case the degree of error made in imitation of other's behaviors. The second interpretation is that of parents and their offspring, where the offspring is taught certain norms of behavior, but not perfectly. In line with Tooby and Cosmides (1990), I assume that natural selection shapes decision rules and the cues they monitor. The degree of mutation can have an effect on the degree of cooperation as shown in Fig. 6. The most sensitive mutation rate is that of symbols. When agents perfectly copy the symbols, this will lead to a reduction in the variety of symbols. As was shown in Fig. 2, a reduction of symbolic diversity makes it more difficult to distinguish trustworthy and non-trustworthy agents. Perfect copies of symbols but not perfect copies of α and β lead to less information available to recognize agent types. On the other hand, when the mutation rate is high, agents are unable to learn which symbols relate to which types of players since they try to learn a secret code that is continuously changing.

The relatively low sensitivity of the not directly observable characteristics of the players is interesting since the results seem to be robust regarding the way agent characteristics are passed through generations. As long as the agents have similar, but not the same, symbols as their parents and have the ability to learn, imperfect transmissions of knowledge (weights) and norms (α and β) between "parent" and "offspring" do not affect long-term cooperation. The result for the weights

Please cite this article as: Marco A. Janssen, Evolution of cooperation in a one-shot Prisoner's Dilemma based on recognition of trustworthy and untrustworthy agents, Journal of Economic Behavior & Organization (2006), doi:10.1016/j.jebo.2006.02.004

is not surprising since the agents update their weights due to learning within a generation. The relative insensitivity for α and β can be explained by the fact that once α is larger than 0.25, the temptation to defect is not valued with the highest utility, and a small mutation of α will not have a significant effect as long as the mutation does not cause α to be lower than 0.25.

6. Conclusions

A model was presented of agents playing one-shot Prisoner's Dilemma games with the option of withdrawal from playing the game when they expect that the opponent is not trustworthy. Over generations, agents evolve to those with social-welfare preferences roughly in line with experimental evidence. The inclusion of the ability to learn to recognize trustworthiness of others leads to levels of cooperation that are significantly higher than would be expected for one-shot Prisoner Dilemma games but are more in line with observations in laboratory experiments with communication.

Based on visual characteristics, agents seem to be able to estimate the trustworthiness of others to such a degree that a high level of cooperation can be maintained, and this result is robust for modest degrees of making errors. In most of my simulations I use a population of 100 agents, and experiments with 1000 agents showed that only a lower level of cooperation can be derived and that more symbols are needed to reach these levels of cooperation. Even when the initial population is dominated by selfish individuals, the evolution drives the model towards agents with a level of other regarding preferences that enables a high level of cooperation.

Although symbols can help to foster cooperation, I expect that specific types of symbols such as reputation symbols, and other forms of information transmission such as gossip, seem to be necessary to derive high levels of cooperation in larger groups. Nevertheless, the use of only visual symbols can explain a possible origin in using symbols as a way to detect cheaters. In this respect, this article contributes to the question why and when humans cooperate in non-repeated social Dilemmas and how this may affect the evolution of symbolic systems that foster cooperation in human societies.

Acknowledgments

The author thanks Daniel Stow for his help with the implementation of the computer program, and T.-K. Ahn, Jerry Busemeyer, Ryan McAllister, Lin Ostrom, Jimmy Walker, and two anonymous reviewers for useful discussions related to this article. Support from the Resilience Alliance, the National Science Foundation (SES0083511), and the European Union (contract nr. IST-2000-26016) is gratefully acknowledged.

References

Ahn, T.-K., Ostrom, E., Schmidt, D., Shupp, R., Walker, J., 2001. Cooperation in PD games: fear, greed, and history of play. Public Choice 106, 137–155.

Ahn, T.K., Ostrom, E., Walker, J., 2003. Incorporating motivational heterogeneity into game theoretic models of collective action. Public Choice 117, 295–314.

Alexander, R.D., 1987. The Biology of Moral Systems. Aldine de Gruyter, New York.

Anderson, S.P., Goeree, J.K., Holt, C.A., in press. Logit equilibrium models of anomalous behavior: What to do when the Nash equilibrium says one thing and the data something else. In: Plott, C., Smith, V., (Eds.), Handbook of Experimental Economic Results. New York: Elsevier.

Please cite this article as: Marco A. Janssen, Evolution of cooperation in a one-shot Prisoner's Dilemma based on recognition of trustworthy and untrustworthy agents, Journal of Economic Behavior & Organization (2006), doi:10.1016/j.jebo.2006.02.004

Andreoni, J., Miller, J., 2002. Giving according to GARP: an experimental test of the consistency of preferences for altruism. Econometrica 70, 737–753.

Ashlock, D., Smucker, M., Stanley, A., Tesfatsion, L., 1996. Preferential partner selection in an evolutionary study of the Prisoner's Dilemma. BioSystems 37, 99–125.

Axelrod, R., 1984. The Evolution of Cooperation. Basic Books, New York.

Berninghaus, S., Güth, W., Kliemt, H., 2003. From teleology to evolution: bridging the gap between rationality and adaptation in social explanation. Journal of Evolutionary Economics 13 (4), 385–410.

Charness, G., Rabin, M., 2002. Understanding social preferences with simple tests. Quarterly Journal of Economics 117 (3), 817–869.

Cosmides, L., 1989. The logic of social exchange: has selection shaped how humans reason? Studies with the Wason selection task. Cognition 31, 187–276.

Fehr, E., Gächter, S., 2002. Altruistic punishment in humans. Nature 415, 137–140.

Field, A.J., 2001. Altruistically Included? The University of Michigan Press, Ann Arbor.

Frank, R.H., 1987. If *Homo Economicus* could choose his own utility function, would he want one with a conscience? American Economic Review 77, 593–604.

Frank, R.H., Gilovich, T., Regan, D., 1993. The evolution of one-shot cooperation: An experiment. Ethology and Sociobiology 14, 247–256.

Gintis, H., Smith, E., Bowles, S., 2001. Costly signaling and cooperation. Journal of Theoretical Biology 213, 103–119.

Güth, W., Kliemt, H., 1998. The indirect evolutionary approach: bridging the gap between rationality and adaptation. Rationality and Society 10, 377–399.

Hales, D., 2001. Tag-based cooperation in artificial societies, Unpublished Ph.D. thesis, Department of Computer Science, University of Essex, Essex, United Kingdom.

Hamilton, W.D., 1964. Genetical evolution of social behavior I and II. Journal of Theoretical Biology 7, 1–52.

Hauk, E., Nagel, R., 2001. Choice of partners in multiple two-person Prisoner's Dilemma games: an experimental study. Journal of Conflict Resolution 45, 770–793.

Janssen, M.A., Ostrom, E., 2006. Adoption of a new regulation for the governance of common-pool resources by a heterogeneous population. In: Baland, J.M., Bardham, P., Bowles, S. (Eds.), Inequality, Cooperation and Environmental Sustainability. Princeton University Press.

Leimar, O., Hammerstein, P., 2001. Evolution of cooperation through indirect reciprocity. Proceedings of the Royal Society London B 268, 745–753.

Lindgren, K., Nordahl, M.G., 1994. Artificial food webs. In: Langton, C.G. (Ed.), Artificial Life III. Reading. Addison-Wesley, MA, pp. 73–104.

Lotem, A., Fishman, M.A., Stone, L., 1999. Evolution of cooperation between individuals. Nature 400, 226–227.

Macy, M., Skvoretz, J., 1998. The evolution of trust and cooperation between strangers: a computational model. American Sociological Review 63, 638–660.

Mealey, L., Daood, C., Krage, M., 1996. Enhanced memory for faces of cheaters. Ethology and Sociobiology 17, 119–128.

Mehrota, K., Mohan, C.K., Ranka, S., 1997. Elements of Artificial Neural Networks. MIT Press, Cambridge, MA.

Miller, J.H., Butt, C.T., Rode, D., 2002. Communication and cooperation. Journal of Economic Behavior and Organization 47 (2), 179–195.

Nowak, M.A., Sigmund, K., 1998. Evolution of indirect reciprocity by image scoring. Nature 393, 573–577.

Oda, R., 1997. Biased face recognition in the Prisoner's Dilemma. Evolution and Human Behavior 18, 309–315.

Orbell, J.M., Dawes, R.M., 1991. Social welfare, cooperators' advantage and the option of not playing the game. American Sociological Review 58, 787–800.

Orbell, J.M., Schwartz-Shea, P., Simmons, R.T., 1984. Do cooperators exit more readily than defectors? American Political Science Review 78, 147–162.

Ostrom, E., Walker, J. (Eds.), 2003. Trust, Reciprocity, and Gains from Association: Interdisciplinary Lessons from Experimental Research. Russell Sage Foundation, New York.

Riolo, R., 1997. The Effects of Tag-Mediated Selection of Partners in Evolving Populations Playing the Iterated Prisoner's Dilemma. Working Paper 97-02-016, Santa Fe, NM: Santa Fe Institute.

Riolo, R.L., Cohen, M.D., Axelrod, R., 2001. Evolution of cooperation without reciprocity. Nature 414, 441–443.

Schluessler, R., 1989. Exit threats and cooperation under anonymity. Journal of Conflict Resolution 33, 728–749.

Skyrms, B., 2002. Signals, evolution and the explanatory power of transient information. Philosophy of Science 69, 407–428.

Stanley, E.A., Ashlock, D., Tesfatsion, L., 1994. Iterated Prisoner's Dilemma with choice and refusal of partners. In: Langton, C.G. (Ed.), Artificial Life III. Addison-Wesley, Reading, MA, pp. 131–176.

Please cite this article as: Marco A. Janssen, Evolution of cooperation in a one-shot Prisoner's Dilemma based on recognition of trustworthy and untrustworthy agents, Journal of Economic Behavior & Organization (2006), doi:10.1016/j.jebo.2006.02.004

Tooby, J., Cosmides, L., 1990. The past explains the present: emotional adaptations and the structure of ancestral environments. Ethology and Sociobiology 11, 375–424.
Trivers, R., 1971. The evolution of reciprocal altruism. Quarterly Review of Biology 46, 35–57.
Tullock, G., 1985. Adam Smith and the Prisoner's Dilemma. Quarterly Journal of Economics 100, 1073–1081.
Vanberg, V.J., Congelton, R.D., 1992. Rationality, morality and exit. American Political Science Review 86, 418–431.
Wedekind, C., Milinski, M., 2000. Cooperation through image scoring in humans. Science 288, 850–852.
Zahavi, A., 1977. The cost of honesty (further remarks on the handicap principle). Journal of Theoretical Biology 67, 603–605.

Please cite this article as: Marco A. Janssen, Evolution of cooperation in a one-shot Prisoner's Dilemma based on recognition of trustworthy and untrustworthy agents, Journal of Economic Behavior & Organization (2006), doi:10.1016/j.jebo.2006.02.004