

Cooperation and Trust in the Presence of Bias *

Murat Şensoy
Department of Computer Science
Ozyegin University, Istanbul, Turkey
murat.sensoy@ozyegin.edu.tr

Abstract

Stereotypes may influence the attitudes that individuals have towards others. Stereotypes, therefore, represent biases toward and against others. In this paper, we formalise stereotypical bias within trust evaluations. Then, using the iterated prisoners' dilemma game, we quantitatively analyse how cooperation and mutual trust between self-interested agents are affected by stereotypical bias. We present two key findings: i) stereotypical bias of one player may inhibit cooperation by creating incentives for others to defect, ii) even if only one of the players has a stereotypical bias, convergence of mutual trust between players may be strictly determined by the bias.

1 Introduction

Stereotypes are beliefs about specific social groups or types of individuals [1]. They play a significant role in our daily interactions with others. Stereotyping is an overgeneralisation of people into groups and categories based on their observable features (e.g., race, gender and so on). Then, these people are evaluated based on the groups they belong to, rather than their own merits. This could be considered useful especially in the absence of personal familiarity with these people [2].

Trust is a fundamental concern in open systems. It lies at the core of all interactions between the entities in uncertain environments. Even though individuals may benefit from cooperation, in certain settings, they may not cooperate in the lack of mutual trust [3, 4]. Most existing models of trust assume that trust between parties builds over time through interactions [5, 6]. On the other hand, stereotypes of the parties may determine their attitude toward each other in these interactions and significantly influence trust and cooperation between them [1]. Hence, there could be an interesting connection between stereotypes and the phenomena like *trust* and *cooperation*.

In this paper, we analyse how stereotypical bias may influence emergence of mutual trust and cooperation between self-interested rational agents. For this purpose, we have used the iterated prisoners' dilemma (*iPD*) game, which is fundamental to certain theories of human cooperation and trust [4]. We have proposed an intuitively way for integrating stereotypical bias into a well-known statistical trust model. This allows us to quantitatively analyse evolution of trust and cooperation between self-interested agents in the presence of stereotypical bias.

By analysing utility functions of agents in *iPD* games, we show that stereotypical bias (either positive or negative) of a player in an *iPD* game creates incentives for the other player to defect. We have also analysed the evolution of trust between players in settings where only one of the players has stereotypical bias and the other is willing to cooperate as much as his opponent cooperates back. We show that, in these settings, stereotypical bias precisely determines how mutual trust converges. This indicates that stereotypes become *self-fulfilling prophecies* in such settings.

Dr. Şensoy thanks to the U.S. Army Research Laboratory for its support under grant W911NF-13-1-0243 and The Scientific and Technological Research Council of Turkey (TUBITAK) for its support under grant 113E238.

Copyright © by the paper's authors. Copying permitted only for private and academic purposes.

In: R. Cohen, R. Falcone and T. J. Norman (eds.): Proceedings of the 17th International Workshop on Trust in Agent Societies, Paris, France, 05-MAY-2014, published at <http://ceur-ws.org>

2 Modelling Trust

We can define *trust* broadly as the willingness of one party (trustor) to rely on the actions of another (trustee) [7]. Several approaches have been proposed to model trust in multi-agent systems. A number of these approaches are based on *Subjective Logic* [8]. Subjective Logic (*SL*) is a belief calculus that allows agents to express opinions as degrees of belief, disbelief and uncertainty about propositions. Let ρ be a proposition such as “agent y is trustworthy in context c ”. Then, the binary opinion of agent x about ρ is equivalent to a Beta distribution. That is, the binomial opinion about the truth of a proposition ρ is represented as the tuple (b, d, u, a) , where b is the belief that ρ is true, d is the belief that ρ is false, u is the uncertainty, and a is the base rate (a priori probability in the absence of evidence), as well as $b + d + u = 1.0$ and $b, d, u, a \in [0, 1]$. Opinions are formed on the basis of positive and negative evidence, possibly aggregated from different sources. Let r and s be the number of positive and negative past observations about y respectively, regarding ρ . Then, b, d , and u are computed based on Equation 1.

$$\begin{aligned} b &= \frac{r}{r + s + w} \\ d &= \frac{s}{r + s + w} \\ u &= \frac{w}{r + s + w} \end{aligned} \quad (1)$$

Based on Equation 1 and inspired by the Subjective Logic [8], the opinion’s probability expectation value is computed using Equation 2. Given the proposition ρ , the computed expectation value can be used by x as the trustworthiness of y in the context c [8].

$$\begin{aligned} t_{y:c}^x(\langle r, s \rangle, \langle a, w \rangle) &= b + a \times u \\ &= \frac{r + a \times w}{r + s + w} \end{aligned} \quad (2)$$

In Equation 2, $w \geq 2$ is a constant representing the non-informative prior weight. The base rate parameter a represents a priori degree of trust x has about y in context c , before any evidence has been received. While x has more evidence to evaluate trustworthiness of y , the uncertainty u (and the effect of a) decreases. In most existing trust models based on *SL*, e.g., Beta Reputation System (BRS) [5] and TRAVOS [9], a and w are set to 0.5 and 2, respectively. Setting $a = 0.5$ means that both outcomes are equally likely if there is no evidence. It should be noted that setting a to 0.5 is in itself a bias. Selecting $w > 2$ will result in evidence having relatively less influence over the trust computations.

3 Modelling Trust with Bias

People use stereotypes to estimate trustworthiness of others in the lack of evidence [1]. However, as they interact with those, they may gather evidence and evaluate their trustworthiness based on it. That is, the bias enforced by stereotypes are replaced by a more objective evaluation based on evidence. For instance, a person with stereotype “*immigrants are not trustworthy*” may consider all immigrants untrustworthy initially. However, some immigrants may be evaluated as trustworthy after a set of pleasant interactions.

Some stereotypes can be stronger than others and direct evidence may have little influence over the bias enforced by them. Therefore, stereotypes about trustworthiness of a group do not only enforce a bias, but also associate a weight with it. Based on this observation, we propose to formalise stereotypes about trustworthiness as in Definition 1. This formalisation allows us naturally incorporate stereotypical bias into trust evaluations based on Equation 2.

Definition 1 A stereotype s is a mapping $\langle g, c \rangle \mapsto \langle a_s, w_s \rangle$, where g is a group, c is context, a_s is the base rate enforced by s in c for members of g , and $w_s \geq 2$ is the weight (or strength) of the stereotype. ■

Let an agent x have a stereotype s formalised as $\langle g, c \rangle \mapsto \langle a_s, w_s \rangle$. If y is a member of the group g , then s enforces $\langle a_s, w_s \rangle$ as the stereotypical bias for y in context c . As a result, x uses the formula for $t_{y:c}^x(\langle r, s \rangle, \langle a_s, w_s \rangle)$ in Equation 2 while evaluating trustworthiness of y in context c . The higher the weigh w_s , the harder it will be to deviate from the base rate a_s as evidence is acquired. That is, if w_s is big enough, then y ’s trustworthiness for x will not depend on the behaviour of y but x ’s stereotype about y . In real life, there are many examples of this extremity. For instance, for a racist, immigrants are untrustworthy in any context no matter how well they behave. This stereotype can be formalised as $\langle \text{immigrant}, \text{any} \rangle \mapsto \langle 0.0, \infty \rangle$, where ∞ is a very big number.

Prisoners' Dilemma Game				Example Pay-off Matrices				
Player 1		Player 2		where	M1		M2	
		C	D		C	D	C	D
C	R	S		T > R > P > S	1	0	1	0
D	T	P		2R > T + S	1.5	0.6	1.5	0.4
					T+S < R+P		T+S > R+P	

Figure 1: Player 1's payoffs for Prisoners' Dilemma.

Here, we should emphasize that existing SL-based trust models such as BRS and TRAVOS implicitly incorporate the bias $\langle 0.5, 2 \rangle$ while modelling trustworthiness.

In the next section, we introduce a game theoretical interaction model which will be used throughout the paper to analyse how stereotypical bias affects cooperation and trust.

4 Prisoners' Dilemma and Trust

Emergence of cooperation among self-interested agents has been widely studied both in computer science and social sciences. In many of these studies, Prisoners' Dilemma (*PD*) has served as a tool to understand and evaluate the dynamics behind strategic behaviour and emergence of cooperation [4]. In its simplest form, the *PD* is a single-shot two-player game where the players have two available actions: *cooperate* (C) and *defect* (D). In this game, the players have the same pay-off matrix. The pay-off matrix and the constraints of the game are shown in Figure 1, where T , R , P , and S are pay-offs. Given the pay-offs, the rational choice of each player (i.e., the Nash equilibrium) is to defect, since D is the dominating strategy for both players. However, this means that the pay-off each player receives is less than it would be if they had both cooperated. This is the so called *dilemma* in this game.

Although defection is the best response in classical single-shot *PD* games, cooperation becomes a social equilibrium in iterated Prisoners' Dilemma (*iPD*) games, where players play the game multiple times and at each iteration, they may determine their actions based on the previous iterations. That is, *iPD* allows players to estimate trustworthiness of their opponents through interaction and determine their actions based on trust [10, 11]. Here, trust of a player in his opponent equates to *how much the player is willing to cooperate taking the risk of defection by his opponent*. Repeated interactions may lead to the emergence of cooperation between self-interested utility-maximising agents [3, 12]. For instance, consider the simple yet effective altruistic strategy *tit-for-tat* in Definition 2. It has been shown that *tit-for-tat* is one of the best strategies and promotes cooperation in *iPD* game [4].

Definition 2 *Tit-for-tat (TFT)* is a well-known strategy in *iPD* game. An agent using this strategy will initially cooperate, then copy his opponent's previous action in the following iterations. If the opponent previously was defective, the agent is defective; otherwise the agent is cooperative. ■

5 Emergence of Cooperation

In this section, we demonstrate under which situations *cooperation* becomes the best strategy in the *iPD* game. Let $0 \leq q \leq 1$ be player 2's level of trust in player 1 and $0 \leq p \leq 1$ be the probability that player 1 selects C at each iteration i of the *iPD* game. Then, the expected utility of player 1 at the iteration can be formalised as in Equation 3.

$$E_i(p, q) = pqR + p(1 - q)S + (1 - p)qT + (1 - p)(1 - q)P \quad (3)$$

Here, for the sake of clarity, we consider scenarios where player 1 determines p before starting an *iPD* game and stick to it throughout the game. Consider an *iPD* game, where q is independent of p ; player 2's trust in player 1 does not depend on his previous actions. Given this assumption, let us analyse an iteration of the game. Equation 4 demonstrates the first and second derivative of $E_i(p, q)$ with respect to p . Given the constraints on pay-offs in Figure 1, the first and second derivatives reveal that the expected utility of player 1 strictly decreases while p increases. Hence, the expected utility is maximised when player 1 does not trust player 2 ($p = 0$), which means defect is strictly the best strategy when p and q are independent.

$$\frac{\partial E_i(p, q)}{\partial p} = q(R - T) + (q - 1)(P - S) < 0 \text{ and } \frac{\partial^2 E_i(p, q)}{\partial p^2} = 0 \quad (4)$$

What would happen if p and q were not independent? Let player 2 accurately estimate p and set q equivalent to p ; that is, player 2 trusts player 1 as much as the probability that player 1 chooses cooperation. Although this assumption

looks very strong at the first glance, *tit-for-tat* corresponds to it when $p, q \in \{0, 1\}$. In this case, the best strategy becomes cooperation for player 1. We can easily prove this intuitive conclusion as follows. Equations 5 and 6 show the first and second derivatives of $E_i(p, p)$ (the expected utility when $p = q$), respectively. The function $E_i(p, p)$ has a extrema (maxima or minima) at the p value defined in Equation 7; let this value be called p^* . This extrema cannot be a maxima for $0 \leq p^* \leq 1$, since the second derivative cannot be negative. That is, the second derivative becomes negative only when $R + P < S + T$. However, then $S + T < 2P < R + P$ has to be true to make $p^* \geq 0$, which leads to a contradiction. As a result, the extrema can only be a minima for $0 \leq p^* \leq 1$. Therefore, to find p maximising $E_i(p, p)$, we need to check only two critical points, i.e. 0 and 1. Given that $P < R$, 1 is the p value maximising $E_i(p, p)$.

$$\frac{\partial E_i(p, p)}{\partial p} = 2p(R + P - S - T) + T + S - 2P \quad (5)$$

$$\frac{\partial^2 E_i(p, p)}{\partial p^2} = 2(R + P - S - T) \quad (6)$$

$$p = \frac{2P - S - T}{2(R + P - S - T)} \text{ then } \frac{\partial E_i(p, p)}{\partial p} = 0 \quad (7)$$

Here, we show that *defection* is the best strategy if players' trust in their opponents is not effected by their actions; e.g., if their opponents unconditionally trust or distrust them, the player's best action would be to defect. We also show that if players' actions precisely determine their opponents' trust in them, *cooperation* becomes the best move.

6 Cooperation and Bias

In this paper, we assume that players in an *iPD* game build an interaction history throughout the game. Now consider a scenario where player 2 has a stereotypical bias $\langle a, w \rangle$ toward player 1 and uses the stereotype-driven strategy in Definition 3 during the game. That is, at each iteration i of an *iPD* game, player 2 uses Equation 2 to compute q_i (the estimated trustworthiness of player 1 at i) and uses q_i to select his move at i . Let us note that even if p is constant, q_i may change as player 2 obtains more evidence about the trustworthiness of player 1 in the game.

Definition 3 Stereotype-driven strategy is used by players with stereotypical bias $\langle a, w \rangle$ toward their opponents. At each iteration, these players use Equation 2 and their interaction history to compute expected trustworthiness of their opponents and select their moves based on it. ■

Player 1 may estimate the best p to use throughout the game as follows. At iteration i , player 2 would have an interaction history in which the *expected number of observations* where player 1 was cooperative would be $i \times p$. That is, expected values of r and s in Equation 1 would be $i \times p$ and $i \times (1 - p)$, respectively. Therefore, the expected value of q_i can be estimated by player 1 using Equation 8.

$$\bar{q}_i = \frac{\bar{r}}{\bar{r} + (i - \bar{r}) + w} + \frac{a \times w}{\bar{r} + (i - \bar{r}) + w} = \frac{i \times p + a \times w}{w + i} \quad (8)$$

Based on q_i , we calculate $E_i(p)$, i.e., the expected utility of player 1 at the iteration i for a given p , using Equation 9. Similarly, $C_i(p)$, i.e., the cumulative utility of player 1 at the end of i^{th} iteration, is computed using Equation 10. Equations 11 and 12 show $E_i(0)$ and $E_i(1)$.

$$E_i(p) = E_i(p, \bar{q}_i) \quad (9)$$

$$C_i(p) = \sum_{j=0}^i E_j(p) \quad (10)$$

$$E_i(0) = \frac{a \times w}{w + i} \times (T - P) + P \quad (11)$$

$$E_i(1) = \frac{a \times w + i}{w + i} \times (R - S) + S \quad (12)$$

In the previous section, we have showed that the best action is to cooperate for player 1 if player 2 can perfectly estimate p and evaluate the trustworthiness of player 1 precisely and correctly without any bias, i.e., $q = p$. On the other hand, in this section, we introduce stereotypical bias in the evaluation of player 2's trust in player 1. In the existence of a stereotypical bias, cooperation may not always yield a better utility for player 1. To understand the effect of stereotypical bias on the utility, we analyse the expected utilities for cooperation and defection at each iteration during the game. Equation 13 indicates at which iteration utility of cooperation exceeds that of defection for player 1. The equation shows that cooperation becomes more profitable for player 1 only at iteration i , which is strictly determined by the stereotypical bias $\langle a, w \rangle$ and

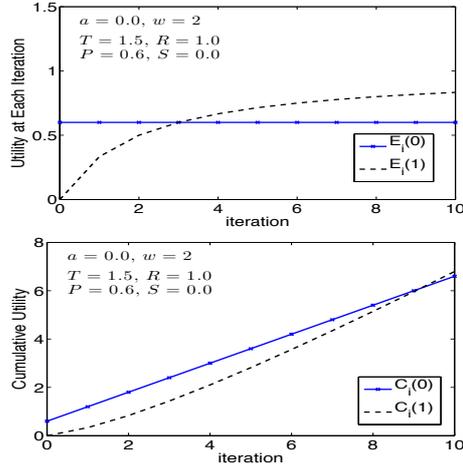


Figure 2: Given $M1$, player 1's utility at each iteration and his cumulative utility when player 2 has bias $\langle 0.0, 2 \rangle$.

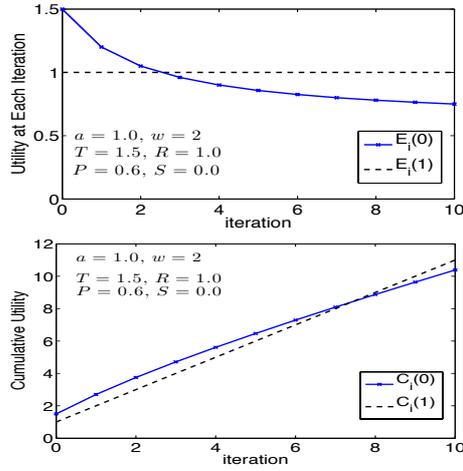


Figure 3: Given $M1$, player 1's utility at each iteration and his cumulative utility when player 2 has bias $\langle 1.0, 2 \rangle$.

monotonically increases with w regardless of a . Let us note that, here we are talking about utility for an individual iteration; cumulative utility of cooperation in the iPD game may exceed that of defection much later.

$$E_i(0) \leq E_i(1) \text{ if } i \geq \frac{a \times w \times (T - R) + w \times (P - S) \times (1 - a)}{R - P} \quad (13)$$

These theoretical results can be supported by empirical analysis using the pay-off matrix $M1$ of Figure 1 where pay-offs are defined as $T = 1.5$, $R = 1.0$, $P = 0.6$, and $S = 0$; therefore, $T + S > R + P$. Let us assume that player 2 has stereotypical bias $\langle 0.0, 2 \rangle$ toward player 1 in an iPD game. Player 1's utilities for $p = 0$ and $p = 1$ are shown in the graphs of Figure 2. These graphs indicate that $E_i(1)$ exceeds $E_i(0)$ after the first 4 iterations, while $C_i(1)$ exceeds $C_i(0)$ after first 10 iterations. This means that if player 1 prefers always being cooperative, he has to wait 10 iterations on average before its cumulative utility exceeds what he would have if he had always preferred *defection*. Therefore, it is more profitable to defect for player 1 if he will interact with player 2 only a limited number of times, e.g., less than 10.

Now, consider the case that player 2 has stereotypical bias $\langle 1.0, 2 \rangle$ toward player 1. Here, player 2 starts the game by completely trusting player 1. Utilities of player 1 is shown in the graphs of Figure 3, where $E_i(0) \leq E_i(1)$ after 4 iterations and $C_i(0) \leq C_i(1)$ after 9 iterations. These results show that both completely positive ($a = 1.0$) and negative ($a = 0.0$) stereotypical bias toward player 1 create some incentive for him to prefer *defection* over *cooperation*, even for the smallest bias weight ($w = 2$).

As w increases, player 1 has a greater incentive to defect no matter what a is. Figures 4 and 5 illustrate the number of iterations necessary for player 1 to start profiting more from cooperation for pay-off matrices $M1$ and $M2$. In $M2$, pay-offs are defined as $T = 1.5$, $R = 1.0$, $P = 0.4$, and $S = 0$; therefore, $T + S < R + P$. The figures imply that if player 2 has stereotypical bias $\langle 0.0, 100 \rangle$, cumulative utility of cooperation exceeds that of defection after 406 and 159 iterations

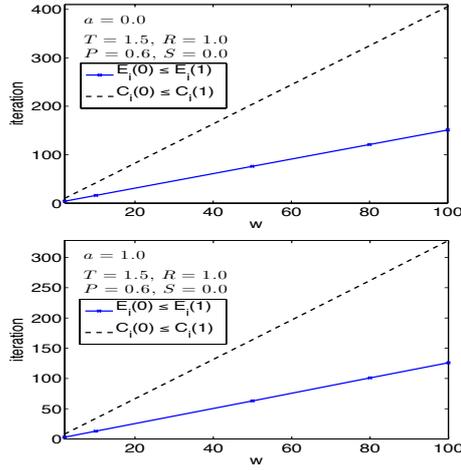


Figure 4: Number of iterations necessary to profit from cooperation for $M1$.

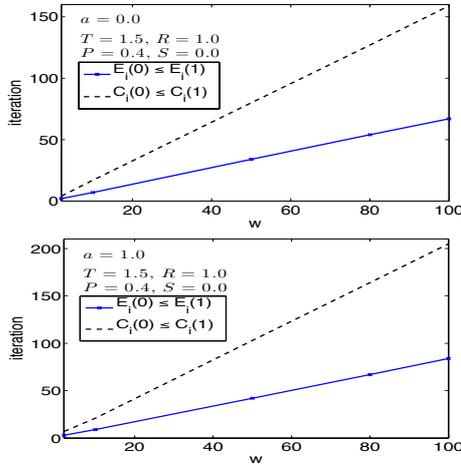


Figure 5: Number of iterations necessary to profit from cooperation for $M2$.

for $M1$ and $M2$, respectively. Similarly, for the bias $\langle 1.0, 100 \rangle$, cooperation is a better choice for player 1 only if he plays the game with player 2 more than 328 and 205 iterations for $M1$ and $M2$, respectively.

In the previous section, we show how cooperation emerges as the best strategy in the iPD game in the absence of stereotypical bias. In this section, we show that stereotypical bias of one player creates incentives for the other to prefer defection over cooperation. Moreover, the incentive for defection increases with the strength of the stereotype. In summary, for an iPD game with finite interactions, defection may become the best strategy for a player if his opponent has a stereotypical bias (either positive or negative).

7 Mutual Trust and Bias

In the previous section, we assume player 1 sets p at the beginning of the game based on its expected utility and does not change it throughout the game. Therefore, over iterations, player 2's trust evaluation in Equation 8 approaches p , i.e., $\lim_{i \rightarrow \infty} \hat{q}_i = p$, unless $w = \infty$. However, in many realistic situations, player 1 may modify p based on the actions of player 2 during the game. Tit-for-tat (TFT) is a well-known example of such a strategy where a player adapts its behaviour based on the past behaviour of his opponent.

In this section, we consider a scenario where player 1 adopts TFT strategy to compute p_i at each iteration while player 2 determines q_i using his interaction history and stereotypical bias as described before. Player 1 using TFT is neither malicious nor unconditionally cooperative. This player starts with cooperation at iteration 0 (i.e., $p_0=1$) and punishes defection with defection; that is, $p_i=0$ if player 2 defects at iteration $i-1$. Even though his opponent has defected previously, TFT quickly forgives and cooperates back if his opponent cooperates in the previous iteration; that is, $p_i=1$ if player 2 cooperates at iteration $i-1$. In this way, TFT reduces losses in utility and promotes cooperation over defection [4].

Table 1: Convergence of mutual trust in iPD games for TFT .

#	Player 1 (TFT)	Player 2 (stereotype-driven)
0	$p_0 = a + \delta = 1$	$q_0 = a$
1	$\bar{p}_1 = \bar{q}_0 = a$	$\bar{q}_1 = \frac{\bar{p}_0 + a \times w}{w+1} = a + \frac{\delta}{w+1}$
2	$\bar{p}_2 = \bar{q}_1 = a + \frac{\delta}{w+1}$	$\bar{q}_2 = \frac{\bar{p}_0 + \bar{p}_1 + a \times w}{w+2} = a + \frac{\delta}{w+2}$
3	$\bar{p}_3 = \bar{q}_2 = a + \frac{\delta}{w+2}$	$\bar{q}_3 = \frac{\bar{p}_0 + \bar{p}_1 + \bar{p}_2 + a \times w}{w+3} = a + \frac{\delta}{w+3} \left(1 + \frac{1}{w+1}\right)$
4	$\bar{p}_4 = \bar{q}_3 = a + \frac{\delta}{w+3} \left(1 + \frac{1}{w+1}\right)$	$\bar{q}_4 = \frac{a \times w + \sum_{i=0}^3 \bar{p}_i}{w+4} = a + \frac{\delta}{w+4} \left(1 + \frac{1}{w+1} + \frac{1}{w+2}\right)$
...
n	$\bar{p}_n = \bar{q}_{n-1} \leq a + \frac{\delta}{w+n-1} \left(1 + \frac{n-3}{w+1}\right)$	$\bar{q}_n = \frac{a \times w + \sum_{i=0}^{n-1} \bar{p}_i}{w+n} \leq a + \frac{\delta}{w+n} \left(1 + \frac{n-2}{w+1}\right)$
►	$\lim_{n \rightarrow \infty} \bar{p}_n \leq a + \frac{\delta}{w+1}$	$\lim_{n \rightarrow \infty} \bar{q}_n \leq a + \frac{\delta}{w+1}$

Table 2: Convergence of mutual trust in iPD games for $pTFT$

#	Player 1 ($pTFT$)	Player 2 (stereotype-driven)
0	$p_0 = a + \delta = 1.0$	$q_0 = a$
1	$\bar{p}_1 = \bar{q}_0 = a$	$\bar{q}_1 = \frac{\bar{p}_0 + a \times w}{w+1} = a + \frac{\delta}{w+1}$
2	$\bar{p}_2 = \frac{\bar{q}_0 + \bar{q}_1}{2} = a + \frac{\delta}{2} \frac{1}{w+1}$	$\bar{q}_2 = \frac{\bar{p}_0 + \bar{p}_1 + a \times w}{w+2} = a + \frac{\delta}{w+2}$
3	$\bar{p}_3 = \frac{\bar{q}_0 + \bar{q}_1 + \bar{q}_2}{3} = a + \frac{\delta}{3} \left(\frac{1}{w+1} + \frac{1}{w+2}\right)$	$\bar{q}_3 = \frac{\bar{p}_0 + \bar{p}_1 + \bar{p}_2 + a \times w}{w+3} = a + \frac{\delta}{w+3} \left(1 + \frac{1}{2(w+1)}\right)$
...
n	$\bar{p}_n = \frac{(n-1) \times \bar{p}_{n-1} + \bar{q}_{n-1}}{n} \leq a + \frac{\delta}{n} \left(\frac{n-1}{w+1}\right)$	$\bar{q}_n = \frac{(w+n-1)\bar{q}_{n-1} + \bar{p}_{n-1}}{w+n} \leq a + \frac{\delta}{w+n} \left(1 + \frac{n-2}{w+1}\right)$
►	$\lim_{n \rightarrow \infty} \bar{p}_n \leq a + \frac{\delta}{w+1}$	$\lim_{n \rightarrow \infty} \bar{q}_n \leq a + \frac{\delta}{w+1}$

In this work, to determine expected outcomes of the interactions between these players in an iPD game, we tabulate expected values of p_i and q_i at each iteration during an iPD game in Table 1. In the table, we reveal an interesting theoretical boundary on the level of emerging mutual trust between players over iterations. Let player 2 have stereotypical bias $\langle a, w \rangle$ toward player 1 and $p_0 = a + \delta = 1$. Then, over iterations, the expected value of both p_i and q_i converge to something less than or equal to $a + \frac{\delta}{w+1}$. For instance, if $a = 0.5$ and $w = 2$, the expected values of p_i and q_i converges something less than 0.6667 and this value further decreases to 0.5455 when $w = 10$. This means that even if player 1 starts with cooperation ($p_0 = 1.0$) and is willing to continue to cooperate as long as player 2 does the same, his probability of selecting cooperation would decrease over time and significantly approach a while w increases. From player 2's point of view, his observations prove that the stereotype provides a good estimate of player 1's trustworthiness, since player 1 behaves almost the same way predicted by the stereotype.

Although TFT is considered one of the best strategies for iPD game, it uses only the previous iteration while making decisions. This property of TFT makes it quick to forget and punish as well. To generalise our findings further, we have introduced *probabilistic tit-for-tat* ($pTFT$) in Definition 4. Similar to the well-known statistical trust approaches [5], $pTFT$ does not consider only the previous interaction but the whole interaction history while computing the probability of cooperation at iteration i (e.g., p_i). Similar to tit-for-tat, $pTFT$ always cooperates unless provoked. However, while tit-for-tat is quick to forgive by considering only the most recent observation, $pTFT$ does not quickly forgive or punish since it does not forget past observations.

Definition 4 Probabilistic tit-for-tat ($pTFT$) is an adaptation of TFT so that it can use the complete interaction history. An agent using this strategy will initially cooperate as in TFT . Then, at iteration i , the agent will cooperate with probability r/i , where r is the number of times his opponent has cooperated in the previous iterations. ■

Table 2 shows that mutual trust between players would be bound by $a + \frac{\delta}{w+1}$ exactly as in Table 1 if player 1 uses $pTFT$ instead of TFT and player 2 uses the stereotype-driven strategy. To empirically demonstrate the convergence of mutual trust, we have simulated 1000 iPD games, each with 100 iterations. In each iPD game, player 1 uses $pTFT$ and player 2 uses the stereotype-driven strategy with $a = 0.5$ and $w = 2$. In Figure 6, we present the mean values of the results collected from the simulations (at top) with their standard deviations (at bottom). Within the first couple of iterations,

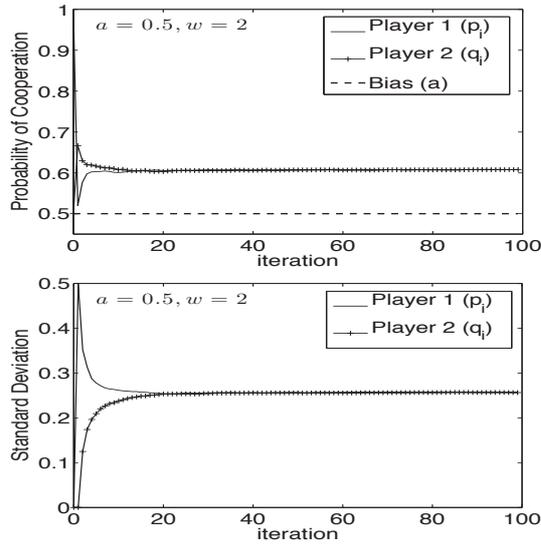


Figure 6: Results for player 1 ($pTFT$) and player 2 (stereotype-driven) with $a = 0.5$ and $w = 2$.

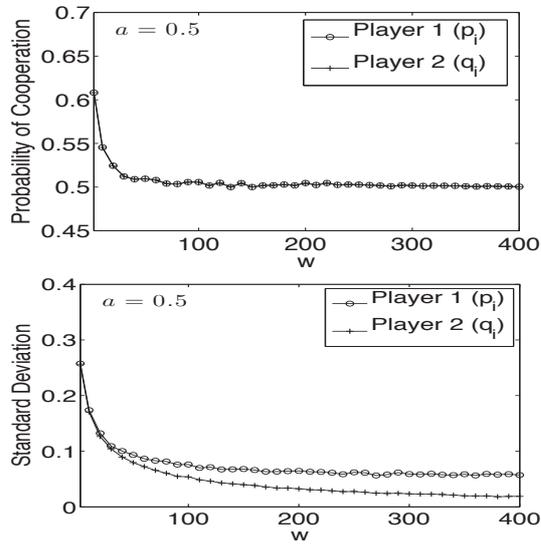


Figure 7: Converged results by w for player 1 ($pTFT$) and player 2 (stereotype-driven) with $a = 0.5$.

mean values of p_i and q_i converge to 0.617 with a standard deviation around 0.25. We have repeated these simulations for various values of w and demonstrated means and standard deviations of our results in Figure 7. These results clearly show that as w increases, mean values of p_i and q_i approach the bias $a = 0.5$ more precisely with a lower standard deviation. When $w = 10$, the means of p_i and q_i converge to 0.539 with a standard deviation of 0.17. However, when w is increased to 220, means of p_i and q_i converge to 0.5002 with standard deviations 0.0626 and 0.0295, respectively.

We also run simulations where both of the players use stereotypical strategy with $a = 0.5$. Our results in Figure 8 demonstrate that both p_i and q_i converge precisely to 0.5 even for the smallest w value (i.e., 2). That is, if both players use state-of-the-art approaches such as *BRS* to model their opponent, their rate of cooperation may not exceed 0.5.

These analyses clearly show that stereotypes become *self-fulfilling prophecies* in the settings where the other party adapts his behaviour based on the behaviour of his opponent as observed in both agent and human societies [4, 13]. Here, we show that if one party has a stereotypical bias toward another, the behaviour of the latter is determined transparently by the stereotype. Furthermore, the behaviour of the other party converges to what is predicted by the stereotype. This justifies the stereotype for its owner, which in turn may increase its strength (i.e., w) and result in a sort of vicious circle. We call this phenomenon *stereotype mirroring*.

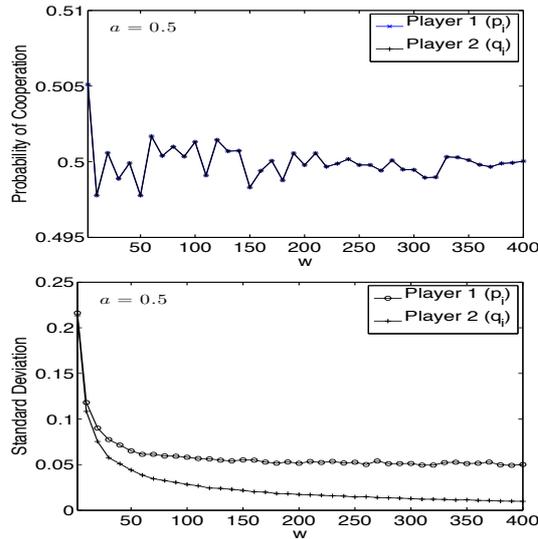


Figure 8: Converged results by w for player 1 (stereotype-driven) and player 2 (stereotype-driven) with $a = 0.5$.

8 Related Work

In this paper, we theoretically and empirically analyse the impacts of stereotypical bias on cooperation and mutual trust. For this purpose, we utilise *iPD* games, which are frequently used in the literature to study evolution of cooperation and mutual trust among self-interested agents.

Axelrod first showed in a computer tournament that strategies based on reciprocal altruism, such as *tit-for-tat*, can lead to the evolution of cooperation in *iPD* game [4]. He discussed in detail that these kinds of strategies foster cooperation and trust even in more complex domains such as international politics. Many examples in human interaction as well as interactions in nature have been modelled and analysed using the *iPD* game in the literature. Cook *et al.* analysed the role of risk taking in building trust relations using a variant of the *iPD* game [13] and conducted extensive experiments on human subjects to examine cross-cultural effects of risk taking on trust building. Bowles and Gintis modelled the economics of ethnic networks composed of agents unified by similarity of one or more ascriptive characteristics [14]. They point out that more efficient communication among the members of these networks creates incentives to maintain them. Brik proposed boosting cooperation by evolving trust between agents in a variant of *iPD* game [3], where trust is represented as a preference to be grouped together with a certain type of agent to play the game.

To model trustworthiness of interacting parties, various cognitive and computational trust models have been proposed in the literature [5]. A number of well-known computational trust models such as BRS [5] and TRAVOS [9] are based on Beta Distributions and correspond to using Subjective Logic with parameters $a = 0.5$ and $w = 2$. This means that an agent using these trust models implicitly incorporates a stereotype $\langle \text{others}, \text{any} \rangle \mapsto \langle 0.5, 2 \rangle$ while evaluating trustworthiness of others. In this paper, we have demonstrated that even such a weak bias may inhibit cooperation and mutual trust in certain settings.

To address bootstrapping problem in trust evaluations, Burnett *et al.* proposed software agents to learn stereotypes based on their prior experience and observable features [15]. Maghami and Sukthankar empirically analysed the impact of stereotyping on task-oriented group formation [16]. They empirically show that learnt stereotypes may restrict agents' social network and group formation abilities significantly in long term. To the best of our knowledge, none of the existing work has formalised stereotypical bias within trust evaluations and quantitatively analysed its effects on cooperation and mutual trust, as we do in this work.

To model trustworthiness of interacting parties, various cognitive and computational trust models have been proposed in the literature [5]. A number of well-known computational trust models such as BRS [17] and TRAVOS [9] are based on Beta Distributions and correspond to using Subjective Logic with parameters $a = 0.5$ and $w = 2$. This means that an agent using these trust models implicitly incorporates a stereotype $\langle \text{others}, \text{any} \rangle \mapsto \langle 0.5, 2 \rangle$ while evaluating trustworthiness of others. In this paper, we have demonstrated that even such a weak stereotypical bias may inhibit cooperation and mutual trust in certain settings.

To address bootstrapping problem in trust evaluations, stereotyping is proposed within multi-agent systems. Liu *et al.* proposed agents to form stereotypes using their previous transactions with others [18]. In their approach, a stereotype contains certain observable features of agents and an expected outcome of the transaction. Burnett *et al.* proposed an agent to generalise its past experience with known agents as stereotypes and use these stereotypes to determine base rate (i.e., a)

while evaluating trustworthiness of new and unknown agents using Subjective Logic [15]. Stereotypes are learned with a regression tree algorithm using a training set composed of observable features of known agents and their trustworthiness. Through simulations, Liu *et al.* and Burnett *et al.* showed that stereotyping is useful to bootstrap trust in certain settings where there is a correlation between behaviour of agents and their observable features.

Maghami and Sukthankar empirically analysed the impact of stereotypes on task-oriented group formation [16]. They equipped agents with simple mechanisms to learn stereotypes based on prior experience and observable features. They empirically show that stereotypes may restrict agents' social network significantly in long term, which in turn may impair the agents' ability to form groups with sufficient diversity of skills. To the best of our knowledge, none of the existing work has formalised stereotypical bias within trust evaluations and quantitatively analysed its effects on cooperation and mutual trust, as we do in this work.

9 Conclusions

In this paper, we have analysed strategies of players in *iPD* game under various settings. First, we have showed that defection is the best strategy if players' trust in their opponents in an *iPD* game is independent of their opponents' actions. However, cooperation becomes the best strategy if players trust in their opponent directly and, without bias, act in a manner that depends upon the actions of opponents. Second, we have analysed the impact of stereotypical bias in *iPD* games and demonstrated that players' positive or negative stereotypes toward their opponents may create incentives for their opponents to prefer defection over cooperation, especially when the number of iterations in the game is bounded. Third, we have showed that stereotypes adopted by one player in an *iPD* game may lead the other player to behave in line with these stereotypes. That is, we have theoretically and empirically analysed how stereotypes become self-fulfilling prophecies in environments where players adapt their behaviour based on the trustworthiness of their opponents, e.g., using strategies like *tit-for-tat*.

In social systems, each agent will have different stereotypes for different groups and interacts with others from different groups over time. It is the mutual experience of agents with slightly different stereotypes and stubbornness of views that makes the evolution of trust interesting. We would like to investigate this phenomenon in the future.

References

- [1] J. L. Hilton and W. von Hippel, "Stereotypes," *Annual Review of Psychology*, vol. 47, no. 1, pp. 237–271, 1996.
- [2] C. Hurst, *Social inequality: Forms, causes, and consequences*. Allyn and Bacon, 1992.
- [3] A. Birk, "Boosting cooperation by evolving trust," *Applied Artificial Intelligence*, vol. 14, no. 8, pp. 769–784, 2000.
- [4] R. M. Axelrod, *The evolution of cooperation*. New York: Basic Books, 1984.
- [5] A. Jøsang, R. Ismail, and C. Boyd, "A survey of trust and reputation systems for online service provision," *Decis. Support Syst.*, vol. 43, pp. 618–644, 2007.
- [6] J. Zhang and R. Cohen, "A framework for trust modeling in multiagent electronic marketplaces with buying advisors to consider varying seller behavior and the limiting of seller bids," *ACM Trans. Intell. Syst. Technol.*, vol. 4, no. 2, pp. 24:1–24:22, 2013.
- [7] R. C. Mayer, J. H. Davis, and F. Schoorman, "An Integrative Model of Organizational Trust," *Academy of Management Review*, vol. 20(3), pp. 709–734, 1995.
- [8] A. Jøsang, *Subjective Logic*, 2011. [Online]. Available: http://folk.uio.no/josang/papers/subjective_logic.pdf
- [9] W. Teacy, J. Patel, N. Jennings, and M. Luck, "TRAVOS: Trust and reputation in the context of inaccurate information sources," *Autonomous Agents and Multi-Agent Systems*, vol. 12, no. 2, pp. 183–198, 2006.
- [10] B. Chen, B. Zhang, and W. Zhu, "Combined trust model based on evidence theory in iterated prisoner's dilemma game," *Systems Science*, vol. 42, no. 1, pp. 63–80, 2011.
- [11] T. Yamagishi, *Trust: The evolutionary game of mind and society*. Springer Verlag, 2011.
- [12] M. Nowak and K. Sigmund, "Evolution of indirect reciprocity," *Nature*, vol. 437, no. 7063, pp. 1291–1298, 2005.

- [13] K. S. Cook, T. Yamagishi, C. Cheshire, R. Cooper, M. Matsuda, and R. Mashima, "Trust building via risk taking: A cross-societal experiment," *Social Psychology Quarterly*, vol. 68, no. 2, pp. 121–142, 2005.
- [14] S. Bowles and H. Gintis, "Persistent parochialism: trust and exclusion in ethnic networks," *Journal of Economic Behavior and Organization*, vol. 55, no. 1, pp. 1–23, 2004.
- [15] C. Burnett, T. J. Norman, and K. Sycara, "Bootstrapping trust evaluations through stereotypes," in *Proceedings of AAMAS*, 2010, pp. 241–248.
- [16] M. Maghami and G. Sukthankar, "An agent-based simulation for investigating the impact of stereotypes on task-oriented group formation," in *Proceedings of Social Computing and Behavioral-Cultural Modeling*, 2011, pp. 252–259.
- [17] A. Jøsang and R. Ismail, "The beta reputation system," in *Proceedings of the Fifteenth Bled Electronic Commerce Conference e-Reality: Constructing the e-Economy*, June 2002, pp. 48–64.
- [18] X. Liu, A. Datta, K. Rzadca, and E.-P. Lim, "Stereotrust: a group based personalized trust model," in *Proceeding of the ACM conference on Information and knowledge management*, 2009, pp. 7–16.