Evolution of Trust and Co-operation

The Prisoner's Dilemma has become widely known and popular model of game theory today. It analyses two participants' decisions concerning co-operation or defection within a structural framework that rewards certain actions and penalises others.[1] The respective consequences of co-operative action or defection are displayed in the matrix below.[2]

Table 1.: Pay-off Matrix in Prisoner's Dilemma

A decision	**B** decision	**A** points received	**B** points received
P (defection)	P (defection)	-2	-2
K (co-operation)	P (defection)	-4	+4
P (defection)	K (co-operation)	+4	-4
K (co-operation)	K (co-operation)	+2	+2

Because of its better application to real life the so-called Iterated Prisoner's Dilemma has become a more important analytical tool. This game does not analyse the consequences of a single game, but rather evaluates the strategies followed by „players" over several „games", as well the results that accrue to each strategy, as assessed by the points they receive in the course of the game.[3] Public attention was drawn to the model by an extraordinary competition held some 25 years ago. Robert Axelrod, a young political scientist, asked acquaintances to develop strategies that might guide players in their decision-making in an iterated version of the Prisoner's Dilemma.[4] Participants in the competition were asked to write down what maxims they would use through several stages of the game. Axelrod converted the mostly verbally formulated strategies into computer programmes and had these programmes compete with each through 200 games.[5] In the game the competing programmes made "decisions" – just as humans would – and received points based on the matrix above depending on the other player's

[1] Larry Samuelson. Evolutionary Games and Equilibrium Selection. The MIT Press. Cambridge, Massachusetts. London, England. 1998. Axelrod, R. and Hamilton, W.D. (1981) Science 211, 1390-1396

[2] The values in the pay-off matrix represent the numbers in the experiment as well as in the games I used myself.

[3] M.A. Nowak, R. M. May (1992) Nature. 359, 826.

[4] Robert Axelrod. The evolution of cooperation. Penguin Books. 1990. London. 2. chapter

[5] Robert Axelrod. The evolution of cooperation. Penguin Books. 1990. London. page 32.

decision. The different programmes accumulated points throughout the games and by the end it became clear which strategy was most successful in maximizing its score.

Later Axelrod made the game more realistic by dropping programmes that reached low scores. Thus in the later stages of the game the more competitive participants remained involved, which made it probable that the „evolutionary competition" would be won by the „fittest" programme(s). The strategy that emerged victorious was formulated by Canadian political scientist Anatol Rapoport and went by the name tit- for- tat (TFT). It consisted of barely two lines: 1) start with C (begin assuming trust), 2) always do as your partner did in the previous game (i.e. mirror her behaviour).

Over the past nearly half a century the Iterated Prisoner's Dilemma has become an extremely popular model in the social sciences. It has been played in manifold variations under varying conditions and with surprisingly different „actors" – humans [6], animals [7], and computers [8]. I have been playing this game regularly for 10 years mostly with college students, but also with corporate customers, and mixed groups. [9] Through these many games I had the chance to observe how participants – using the old established trial and error method – seek and often find the strategy for successful behaviour. Within the normal confines of the game results (scores achieved) were the best basis for comparison. To expand the horizon and possibilities of comparison I conducted a carefully planned series of experiments with the goal of answering the following conditions in a controlled environment:

- Do participants have an identifiable strategy?

[6] M.A. Nowak, K. Sigmund. (1998) Nature. 393, 573

[7] L.A. Dugatkin, Cooperation Among Animals: An Evolutionary Perspective Oxford Univ. Press Princeton, NJ, 1997. és D,.W. Stephens, C. M. McLinn, J. R. Stevens. Discounting and Reciprocity in an Ittereted Prisoners Dilemma. Science Vol. 298. 2002 dec 13. 2216, also here: M. Mesterton- Gibbons és Eldridge S. Adams. The Economics of Animal Cooperation. 2146-2147.

[8] György Szabó és Csaba Tőke. Evolutionary prisoners's dilemma game on a square Lattice. Physical Review. E. Vol. 58. Num.1 Jul. 1998. 69. oldal

[9] Marosán György. A simogatások játékelmélete, avagy a kifizetési mátrix pszichológiája. (A Game Theory of Patting, or the Psychology of Payoff Matrices) Vezetéstudomány 1996. volume 12. page 46. and Reflections at the end of the Millenium. Villányi úti könyvek series. Page 176

- Do the decisions they make in the course of the game reflect their strategy?
- Are participants willing, and/or able to change their starting strategies based on their experiences in the games?
- Which strategy becomes the winner (most successful)?
- Do participants draw correct conclusions from success or failure?

The Experiment's Description

The participants of the experiment were 68 students (39 girls, 29 boys) of the BGF Külkereskedelmi Főiskola (School of Foreign Trade), the Általános Vállalkozási Főiskola (General Enterprise School), and Zsigmond Király Főiskola (King Sigmund College), in groups of 12-16 members. These youth, between 18-22, came of age following the regime transition, and their attitudes differ significantly from the elder generation. Their behaviour was largely shaped in an environment marked by market economy and political democracy. Participants were put in pairs and had to make decisions through 30 games. At each stage the players had the option – based on their general attitudes and the experiences acquired during the game, and with a view towards the pay-off matrix - of noting a "C" for co-operation or a "D" for defection on their sheet of paper. They then received and accumulated points based on the payoff resulting from the decision they and their partners made. The declared goal of the game and each participant was: „accumulate as many points as possible".

Throughout the game – as in every game I organised over the past years – all decisions, as well as the points received in each round of the game and overall, were on display for all to see. We assumed that the overall „population" of participants employed a mix of various strategies and that even individual player's behaviour could often be characterised as a blend of different strategies. The game was played under „noisy" conditions, meaning that decisions were influenced by ambivalent and chance factors.[10] These

[10] Robert Axelrod evaluates the role of and the possibilities for handling a noisy environment. Robert Axelrod. The Complexity of cooperation, Princeton University Press. Princeton, New

factors served to make the game more reflective of real life conditions. The arrangements described above made it possible to observe tendencies in the overall distribution of strategies, as well as any overall shifts in these tendencies throughout the game.

In the first phase – rounds 1-10 – the partners had the opportunity to discuss their respective decisions in round 4, 6 and 9, and to negotiate future terms of partnership. We made it clear that „communication is an option, but is not obligatory. You can make any agreements, but you do not have to keep to them." In the second phase (rounds 10-20), players were permitted to „change partners" following rounds 12, 15 and 18. Any player signalling his or her desire to separate was placed on a list of „free agents". Players on the list received new partners based on a lot. Those who wished could choose negotiation instead of „separation". In the third phase (rounds 20-30) evolution began. Players with the lowest scores were left out of the game after rounds 22, 24 and 28.[11] Players who lost their partners in this process were assigned a new one by lot and the game continued through round 30. Throughout the game we also kept increasing the pay-offs: 2 points (or 4) in the first phase were worth 3 (or 6) in the second phase, and then became 4 (8) in the final phase.

An analysis of the participants' attitudes

Pairs were matched by lot and once they had been teamed they filled out the starting questionnaire prior to the first phase of the game. The rather simple questionnaire asked about the key values underlying participants' attitudes and sought to explore their relationship to their partner.[12] Everyday thinking distinguishes between two fundamentally opposite attitudes: a co-operative disposition based on trust, and a cheating (or defecting, as we use it here),

Jersey. 1997.

[11] Initially I had planned to have more selection, but due to the practical difficulties in organisation I finally decided to have only three selection rounds in the 12 person groups (following rounds 22, 25 and 29), and four in the group of 14 and 16 players (in rounds 22, 23, 26 and 28). Therefore, as two players were removed in each selection round (and in one instance four) altogether 36 players were eliminated from the game.

[12] The values under investigation were trust, willingness to co-operate, orientation towards success, competitiveness, the willingness to exploit others, taking others' interest into consideration, justice and mutuality.

attitude based on distrust and unfriendly competition. Typical life situations on the other hand display a much larger variety of complex and different behavioural patterns. Lately games with human and computer programme participants also partly brought out more nuanced strategies than the crude ones described above. [13] Such strategies include the „Pavlovian" strategy (continue as long as you are successful, but change immediately when you experience failure), or the „forgiving" TFT (do not immediately respond with defection when you are cheated, but do follow the TFT in general), the „distrustful defector", (cheat and abuse the other and then move on while you can), etc. A most recent experiment identified three essentially distinct attitudes: co-operation, guarded co-operation, and competitive non-co-operation. [14]

Based on the analysis of the 68 questionnaires I found that the participants could be grouped into these roughly typical attitudes [15]:

1. „Do as you would be done by" vagy "To err is human; to forgive is divine" (Forgiving TFT – FTFT)

   Advancing trust, unconditional co-operation and assuming that the partner's defection is due to error rather than intent, thus tolerating it over several rounds. Looking out for others and regarding the community as valuable and worth investing into.

[13] Bruno Beaufils, Jean-Paul Delahaye, Philipe Mathieu. Out meeting with Gradual: a Good strategy for the iterated Prisoner's Dilemma. Artifical Life V. Proceedings of the Fifth International Workshop on the Synthesis and Simulation of Living Systems. Edited by Christopher G. Langton, Katsunori Shimohara. A Bradford Book. The MIT Press. Cambridge Massachusetts. London, England. Page 202.

[14] Robert Kurzban – Daniel Houser. Experiments investigating cooperative types in humans. PNAS 2005 Febr. 1. vol 102. no 5. 1803-1807

[15] In the case of 8 values the questionnaire offers three choices for each value, which are "full trust", "conditional co-operation" and "competitive defection". Respondents were asked to distribute 10 points among these three potential answers. During the evaluation of the questionnaires I multiplied the score awarded to each choice by 1, 2, or 3, depending on the given question's degree of "good faith", "distrust" or "selfishness". By adding up these scores the player received an overall score on the scale ranging from the theoretical "Mother Theresa" minimum of 80 points, to the maximum, absolutely selfish "Gordon Gekko" ideal-type of 240 points. Among participants the minimum score was 108 points, while the maximum went up to 218. The scale I used went from 100-220 points, which I divided into units of 6 points to get a scale divided into 20 parts. Through this method I could observe and compare the measured distribution of attitudes on the "trust-competition" scale. Changes in this distribution and the corresponding attitudes also became more easily discernible.

2. „Fool me once, shame on you. Fool me twice, shame on me.” (tit- for- tat - TFT)

   Eager to secure his or her own interest, but fundamentally accepting of the concept of mutual benefit. Conditional co- operation, advancing trust with a strong dose of scepticism and immediate withdrawal of co-operation upon disappointment of said trust. Contributes to the community and respects its benefits.

3. „Put your faith in God, but keep your powder dry.” (distrustful TFT - DTFT)

   Distrustful and risk- minimising (often starting with D), unsure about the partner’s expected behaviour, co- operating conditionally and demanding mutuality. Guided by desire to enforce short- term interest, displays low levels of solidarity, and withdraws from cooperation very easily.

4. „In the end, a man's motives are second to his accomplishments.” (Competitive Defector - CD)

   Regards life as a competition and plays to win. Considers defection (breaking agreements) an acceptable price for victory, but will co-operate if she perceives it to be in her interest. Interprets rules according to her own interests, exploits partners and the community whenever possible, but punishes defection of others immediately.[16]

I assumed that unlike computer programmes human subjects participating in the experiment would not consistently follow a clearly delineated behavioural strategy. Participants in the game would act based on a strategy mix composed of the impressions from the situation at hand, past experience and the partner’s behaviour. Environmental noise consisting of misunderstandings, misinterpretations, or even mishearing, as well as

[16] The different ways in which the various strategies affected the participants’ behaviour can be seen in Table 5, which we compiled based on their behaviour in the first 10 rounds of the game. The numbers presented in the table suggest that the four different strategy types lead to different behaviours, but the divergence cannot be exactly specified.

subconscious decision-making processes („I don't know why I did that") would exert a substantial influence on their choices.

As players could see each other and even interact through negotiations and thereby witness first-hand how agreements were disregarded, the participants' actual behaviour frequently differed from the responses they provided in the questionnaire.[17] In the heat of the game they often strayed from their imagined (or chosen) roles, or they began to behave as „ordinary" people. On the positive side these factors make the game more realistic than computer simulations, but at the same time they also make the results less amenable to quantification, as well as harder to track, two factors that combine to make an interpretation of the experiment's outcome more difficult.

Results

Based on the analysis of the submitted questionnaires Table 2 shows the distribution of basic behavioural strategies chosen by the participants at the beginning and the end of the game.

Table 2: Behavioural strategies selected by participants at the beginning and the end of the experiment

Chosen behavioural strategy	Followers based on starting questionnaire	Followers based on final questionnaire
Forgiving Tit-For-Tat (FTFT)	5 (7.4%)	3 (4.4%)
Tit-For-Tat (TFT)	15 (22.1%)	26 (38.2%)
Distrustful Tit-For-Tat (DTFT)	27 (39,7%)	28 (41.2%)
Competitive Defector (CD)	21 (30.8%)	11 (16.1%)

As a result of the game the participants' opinion tended to gravitate towards the TFT. Graph 1 shows how the distribution of preferences changed as a result of the game. The mean of the starting questionnaire is 12.5 its deviation is 4.2, while the final questionnaire has a mean of 11.06 and a deviation of 3.98.

[17] The effect of reputation was demonstrated by several analyses: Brooks King-Casas, Damon Tomlin, Cedric Anen, Colin F. Camerer, Steven R. Quartz, P. Read Montague. Getting to Know You. Science Vol. 308 2005 - April 1, p. 78

The number of „extreme" opinions changed, and so did the spread of views. The decline in deviation shows that the extreme strategies – that is players either granting unconditional trust or constantly engaging in selfish defection – gradually became relegated to the background. The result of the competition between the various strategies is summarised by Table 3. Based on the evaluation of the questionnaires this table shows the distribution and change in behavioural strategies among the top 10 and top 20 players, as well as those 32 that were not eliminated. These numbers reveal the „evolutionary success" of different behavioural strategies.

Table 3: Changes in behavioural strategies among top performers

	First Ten		First 20		First 32	
	Starting questionnaire	Final questionnaire	Starting questionnaire	Final questionnaire	Starting questionnaire	Final questionnaire
FTFT	2	2	2	2	3	2
TFT	3	5	7	8	11	14
DTFT	4	3	8	8	10	13
CD	1	0	3	2	9	5

The table shows – though not as clearly as Axelrod's competition did[18] – that the TFT-type strategies emerges as „winners" from the competition. The majority recognized and proved the success of TFT-type behavioural strategies.

Success and its measures

Traditionally the success of behavioural strategies is measured by their placement in the competition.[19] In reality it is wise to employ various criteria that show different aspects of success. First and naturally comes the number of winners.[20] Another relevant criterion is – as the medal and score tables during the Olympic games – the average ranking of players using one strategy or the other, in other words the average ranking of a given strategy.[21] Key

[18] In Axelrod's competition there were hardly any "defecting" programmes among the top performers.
[19] R. Axelrod. Evolution of cooperation
[20] Number of winners = Players who made into the first 10.
[21] Average ranking = number of persons following this given strategy among the final 32/32

information is provided by calculating the average score of a given strategy.[22] This is a good reflection of a strategy's success, as it is a game's declared goal to maximize the score. Beyond the above another interesting aspect is the given strategy's „retaining power" and „attractiveness."[23] The results of the game based on the criteria above are as follows:

Table 4: Success measures of different behavioural strategies

Measure of success	FTFT	TFT	DTFT	CD
Winners in the top ten (based on s = starting, or f =final questionnaire)	2 (s), 2 (f)	3 (s) 5 (f)	4 (s) 3 (f)	1 (s) 0 (f)
Average placement (last 32in the game/number of followers)	0.4	0.9	0.5	0.24
Average score achieved	84	92	80	76
Retaining power				
Attractiveness				

When interpreting the results we must take into consideration that in the course of the game more than half the players dropped out, which is also a reflection of their chosen strategies success. This naturally reduced the FTFTs, CDs and DTFTs average scores and ranking. As the selection reflected the strategies' „fitness" and reduced the „population" based on comparative fitness, the selection enhances the conclusions, however. All success criteria point to the success of the TFT strategy.

A detailed analysis of the results – based on tracking the score and negotiations during the first phase of the game and comparing it to the players' later results – paints an interesting picture of the top players' views.[24] During the first round 44 players out of 68 chose to start with a D, and only 24 opted for co-operation. Among the players who ended up in the top 20 the ratio was almost reversed: only 7 began with a defection while 13 co-operated. Following the first negotiation – during which almost all players agreed to

[22] Average score = the overall score achieved by players following a given strategy/ the number of followers. This number reflects a given strategy's fitness.
[23] Retaining power = number of followers at the beginning of the game(based on starting questionnaire)/number of those sticking by the given strategy, and
Attractiveness = number of followers at the beginning of the game (based on starting questionnaire number of followers at the end of the game(based on final questionnaire).
[24] In my opinion the individual administrative sheet only provide basis for analysing the results of the first ten rounds. As the game progressed participants increasingly changed their original strategies, they tired and this makes an interpretation of their behaviour more difficult.

negotiate [25] – 51 participants wrote C, but in the round next only 35 stood by co-operation. Among the players who finished in the top 20 18 opted for C immediately following the negotiation, and even during the next three rounds this number did not fall below 16.

Among those players that classified themselves as DCs only one player made it into the top ten (but in reality he was forced to follow a DTFT behaviour as well). The members of the FTFT pair that made into the final ten – a pair that selected itself, as it emerged from the final questionnaire – steadfastly held out in their commitment to co-operative behaviour. They claimed that they would have given the C for co-operation 30 rounds in advance. At the same time the other three FTFT players were not quite as lucky: they were matched with DCs or DTFTs and quickly fell irretrievably behind and were forced to change strategies.

The transformations brought about by the game are displayed by the transformational matrix in Table 5. This table show the responses based on the starting questionnaire, the details of the modifications created by the game, as well as the results according to the final questionnaire.

Table 5: The transfer matrix

Final Starting	FTFT (3) (receives)	TFT (26) (receives)	DTFT (28) (receives)	CD(11) (receives)
FTFT (5) (gives)	**2**	2	1	0
TFT (15) (gives)	1	**9**	5	0
DTFT (27) (gives)	0	11	**14**	2
CD (21) (gives)	0	4	8	**9**

During the game the number of TFT adherents grew significantly, and the main sources for growth were those who started out with DTFT or CD strategies. Thus the game proved – in contrast to everyday logic and the expectations of most participants, but consistently with Axelrod's results – that the majority of players recognized the TFT-type strategy's effectiveness. This shows that under conditions that best reflect real life – that is an insecure

[25] The negotiations sometimes led to rather curious arrangements. One participant would ask his partner to sacrifice points, for instance, arguing that the other was way ahead, due to the fact that he had started off with Ds. There the arrangement was: I'll write a D now, you write C, and from then on we will both write Cs.

situation, an unknown partner, high risk of being cheated and considerable noise influencing the game – the TFT-type of behaviour is most successful.

The number of FTFT followers decreased. Those among them (and this was the majority, three persons), whose experiences were largely negative learned the following: as long as the majority of your potential partners is distrustful, competitive and tries to exploit her environment, you can only follow a strategy based on unconditional trust with a "tried and tested" partner. The life model provided by the FTFT strategy, which regards defection as a forgivable blunder and thus tolerates it over several rounds, and often employs the principle of „turn the other cheek", is disadvantageous under real life conditions. The expectation that a series of co-operative moves will "convert" the defector is – based on the results of the experiment – unrealistic. Continuous co-operation will rather reinforce the defector in her strategy than induce her to change.

Nonetheless, the two persons for whom this strategy worked felt strengthened in their conviction that the FTFT was a useful guide for behaviour. „If you found the right partner", one of them argued, „you should not try to cheat, or to look for a new partner, but be satisfied and reinforce the good relationship again and again." This perspective underlines that in real life there is no strategy that works in all circumstances. Our behaviour – in accordance with the Contingency Theory of Management – needs to be adjusted to the given environment and, most importantly, to the strategy used by our respective partner.

TFT is not a „winner" in the traditional sense, not the all-time victor of pair competitions. In A. Rapoport's interpretation TFT's success meant that „nice guys sometimes finish first."[26]

The number of DTFTs hardly changed. Some of them became TFTs, led by the insight that it is more advisable to establish good long-term relations and therefore to advance some trust initially, to be more co-operative and less prone to defections in general. A "well-chosen" D – as in the proverbial stick – sends a message that says "get a grip, co-operation is in your interest as well."

[26] Elliott Sober, David Sloan Wilson. Unto Others. Harvard University Press. 1999. London England. Page 86.

[27] The DTFTs who turned into TFTs were replaced by those DCs who in turn switched to DTFT camp. These were players who acknowledged in the course of the game that mutuality and the respect for rules are essential conditions of co-operation and therefore also a condition for maximising one's score in the long-run.

The number of DCs decreased significantly. Many of them decided to change strategy themselves, observing their own results as well as those of the other players. In addition to the lot of „Ds" they were accumulating as a consequence of their defections, negotiations between rounds played a key role in their conversions. During the negotiations players often told each other what they expected the other to do, and making clear what their partner could gain from co-operating or lose by rejecting an agreement or defecting from its terms. The DC group's average behavioural model – see Table 6 – proved to be a failure in the long-run, under the conditions of this game. More than half the DCs switched strategies, therefore, and they mostly joined the DTFTs, though some became TFTs. Even later most of them were continuously preoccupied with the perennial question of "where did I make a mistake, what trick could I have used to score some more points?"

The data clearly shows how the efficient norms enabling community life are enhanced. The use of the word "efficient" is justified by the fact that – as we will see – TFT generates the highest individual and team scores. This could be interpreted to mean that the wealth of communities that consist of individuals following this strategy increases faster than those of communities comprising DCs or DTFTs. The emergence of stable norms is supported – among other things – by the observation that shows a decrease in the variability of behaviour towards the end of the game, which suggests that players attitudes became more uniform towards the end. The fact that the majority opted for a TFT-type of behaviour can be interpreted to mean that the TFT behavioural strategy is selected in the process of the game and

[27] Axelrod, based on Rapoport, already pointed out such an interpretation of a D action taken by a TFT. He also noticed that TFT usually did not emerge victorious in pair competitions. When encountering defection and responding in kind, TFT was not vindictive; it did not respond to a defection with a series of defections, but rather with a warning, as if trying to say: come to your senses, co-operation is better for you, too. Thereby TFTs encourage their partners to co-operate and ensured that it became worthwhile to exchange "happiness" in the long-run.

The fact that the majority opted for a TFT-type of behaviour can be interpreted to mean that this strategy gets selected as a result of the game and rises above the others in a community of people who are suspicious of each other, selfish and prone to defection. With some exaggeration it could be said that the Wall Street's Gordon Gekko is forced to adopt some aspects of Mother Theresa-type of behaviour, while on the other hand the Mother Theresas become alerted to the importance of paying attention to their self-interest. Even without any intervention from above – sans divinely inspired and legitimated ethical principles - gradually a behavioural strategy based on co-operation and advancing trust emerged. [28]

Responses to the questions asked at the beginning of the study

1. Do the participants have a starting strategy?

The starting and final questionnaires clearly showed that the participants had some sort of strategy, though it was never clear-cut, but rather a mix of similar strategy variations. The existence of strategy-mixes reinforces the notion that values are often obscure when it comes to real-life application, and they often do not provide an unambiguous guide to behaviour in concrete situations. At the same time it also means that they are flexible, and they can serve as foundations for different decisions applied to various situations in which the optimal responses differ. Nonetheless, we could identify four different types of attitudes, which also manifested itself in how players chose to interact with their partners. These four types were FTFT, TFT, DTFT and DC attitudes.

2. Are these strategies reflected in the decisions players took during the game?

[28] Jonathan Bendor, Pietr Swistak. The evolution of Norms. Americal Journal of Sociology. Vol. 106. Number 6. 2001 May 1493-1545.

The results suggest that the behavioural strategies indicated in the questionnaires exert an influence on the actual decisions taken by players. The concrete effects could be best observed in the players' responses to unexpected, often surprising moves, or the unanticipated development of the conditions and the results. In the table below we summarise part of the results that emerged from the game's first phase, as well as observations noted by hand in the course of the game.[29] At this point there was still a rather strong connection between the starting strategy chosen by players and their reactions to specific events. Here are some of the typical issues that these strategies had to help resolve: what is the first move (C or D), what is the answer to a partner's C, how do I react to a D, do I negotiate, and do I keep to the agreement I just reached?

Table 6: Comparison of behavioural aspects of different strategies

Type of Strategy	First move	Response to a C (willingness to co-operate)	Response to a D (willingness to defect)	Willingness to negotiate	Keeping agreements initially	Keeping agreements later
FTFT	Always C	Always C	Several Cs (at least 2)	Always	Always	Always
TFT	75% C	80%	Occasional C, mostly D	Always	Mostly	Mostly
DTFT	50% C	70% C	Always D	Sometimes no	Defects often	Usually defects
DC	20% C	45% C (regular defector)	Always D	Often no	Defects mostly	Always defects

[29] These administrative tables have to be interpreted carefully, since the conditions of the game kept changing constantly, therefore they may not be completely comparable. 1) the points the players could receive were increased in the course of the game, which also increased the risk 2) there was the possibility of negotiation, and therefore learning 3) the possibility of separation, and thus a tool for pressuring the partner also became available 4) evolution began to kick in, which increased competition 5) the winning strategy gradually become clearer in the course of the game, which also enhanced the learning process 6) in the later stages the players became tired, they were bored and paid less attention. As a consequence the number of misunderstandings grew which significantly – but in an unpredictable fashion – influenced decision-making. This is why Graph 4 shows only rough percentages.

These data demonstrate that the different strategies lead to distinct pattern of reaction in the various turns of the game, even if they are always consistently applied on an individual level.

3. Do individuals learn from the experience acquired during the game?

The necessity to collect points on the one hand, and the negotiation process as well as the demonstration effect motivate players to rethink their original ideas and sometimes to change them. The DCs, for instance, were forced to give up their original attitude, - or to bow to the inevitable necessities – and to become co-operative. To be more precise, as a group they went through two distinct processes of change, as revealed by the final questionnaires: one part of the group leaned that switching to TFT or DTFT was a better way to go, while another part, making up almost half the group, felt its DC attitude reaffirmed and kept asking herself: where did I make a mistake?

4. Which is the "winning" (achieving the highest score) strategy?

Given our effort to make the environmental conditions as realistic as possible and the small number of participants I did not assume that any single strategy would emerge as the exclusive winner. What I did expect, however – and what did in fact emerge – is that a relatively clear picture would emerge. Graph 3 clearly demonstrates that in the course of the game DCs increasingly abandon their initial strategy and adopt DTFT or TFT strategies. The TFT-type strategies' victory was not as overwhelming as in Axelrod's competition, but given the real life constraints in this experiment it was still impressive enough. Table 6 also underlines this interpretation of the results. At the same time – and this cannot be stressed sufficiently – success is environment and partner dependent.

5. Do participants draw the right conclusions from either success or failure?

Analysis of the starting and the final questionnaire suggests that opinions concerning a "life-strategy" undergo significant changes. On the one hand, Table 1 shows that on the scale from FTFT to DC opinions pull towards the

"centre". This means that the extremely trusting (Mother Theresa- type) and the extremely selfish- competitive (Gordon Gekko- type) attitudes move towards a behaviour that is better characterised by TFT mixed with some distrust. The two extreme strategic approaches were the real losers of this game. The game revealed that neither unconditional trust nor a constant defection without appreciation for the effects of punishment is an efficient strategy under today's conditions, vis- à- vis an average partner. Thus the average player's opinion and behaviour gradually gravitated towards a higher level of trust and co- operation during the course and as a consequence of the game.

It is interesting to observe the pattern of the distribution of the different strategies' respective frequencies. Though we want to avoid drawing far-reaching conclusions from this distribution due to the small size of the sample, it is easy to see that the starting distribution is tilted more towards the competitive strategies, while the final questionnaire reveals a shift towards the Forgiving TFT end. This is also supported by the numerical analysis of the changes in the overall opinions (see Graph 4), which shows that DC opinions switch to DTFT and TFT.

Individual success and communal well- being

Analysing the results of the individual competitions and the pair competitions (the overall scores achieved by pairs) allows for interesting conclusions. The individual competition was won by a DTFT player (based on the starting questionnaire), with 124 points. He was paired with an FTFT in the beginning and acquired a substantial lead through a series of defections, but gradually shifted towards a TFT behavioural mode and ended up adopting that strategy in the final questionnaire. He was followed by two TFTs with 122 points each. Then came two FTFT players with a 120 points each. In the top ten there was also another TFT (112), as well as three DTFTs with 116, 108 and 106 points respectively. A DC player who had also accumulated a substantial score in the first phase, but was then forced to switch to a TFT strategy after his partner punished him with a series of defections which cost him his advantageous

position, also came near the top crowd with 108 points (he regained some of his early advantage with the TFT strategy and also ended up opting for this strategy in the final questionnaire). A look at the final questionnaire (see Table 4) thus shows an even clearer advantage for the TFT strategy: 3 TFTs were at the top, and another two were in the top ten, meaning that half the players in the first ten identified with this strategy.

In the context of the experiment – which attempted to imitate real life conditions – TFT proved to be the best individual strategy. The insight gained from experiments conducted by others shows that in an environment without fixed rules, with participants that are motivated by self-interest only, and a free choice of partners – thus in conditions resembling a Hobbesian war of all against all – the "Pavlovian" strategy (continue a certain behaviour as long as it is successful, but change as soon as it fails) is the most efficient.[30] But as our own experiment showed, from such an environment a population gradually emerges in which DTFT is the most common at first, and then TFT. This demonstrates the success of the TFT strategy.[31] At the same time we would not overstate the success of the FTFT pair by asserting that in a humane environment, where trust is abundant, co-operation normal and noise low, FTFT would be the most successful strategy.

Comparing the joint scores of a pair of players is interesting, because it is relatively easy to score highly if an individual chooses a "cheating" strategy at the other's expense. If this happens, however, the pair overall will accumulate a low score, since the partner will gather negative points in the process. If both partners in a pair have similarly high scores, then that suggests they got "rich" by co-operation. The 32 people remaining standing at the end of the game formed 16 pairs. Of these 16 pairs 9 had been unchanged since the beginning of the game, the other 7 were randomly matched from partners in pairs which had either requested "separation" or in which one partner had performed too weakly to stay in the game. Of the nine pairs surviving until the end there were 1 FTFT-FTFT, 3 TFT-TFT, 2 TFT-DTFT, 2 DTFT-DTFT, and one TFT-CD partnerships (based on the final questionnaire).

[30] Nowak, M., Sigmund, K., 1993. A strategy of win-stay. Lose-shift that outperforms tit-for-tat in the prisoner's dilemma game. Nature 364, 56-58.
[31] Szabó György. A Jó, a Rossz, és a Magányos számítógépes küzdelme. Természettudományi Közlöny. Vol.134/5. P.197.

The competition of pairs was won by the FTFT partnership with 240 points. Among the top 5 pairs of the 9 that had not changed throughout the game, there were also were 2 TFT (224 and 208 points, respectively), 1 DTFT (210) and one TFT-DTFT (206) pairs. This suggests that under the given conditions TFT has the greatest capacity to general communal wealth.

The "community hero"

It was during analysing the data that another fascinating question emerged: which strategy does the most for the community well-being? Or, to put in the game's context, which one maximises a pair's overall score? Strategies have different ways of making participants modify their behaviour. First, the scores displayed on the blackboard for all to see contribute to a demonstration effect. Players can observe how different strategies lead to higher or lower scores. Second, players can influence each other during negotiations, where they can exert pressure either by threatening defection or even separation. Finally, a given strategy will either reward or penalise the given player who follows it, thereby urging her to rethink her strategy from move to move.

When awarding the honour of "community hero" we take two criteria into consideration – without losing sight of the fact that both, individual success and the ability to contribute to community well-being, depend on the given environment and partner. First, we consider the given strategy's success achieved in the competition of pairs. As we saw, this competition was won by the FTFT strategy, though among the top teams in the competition of pairs the TFTs were in a majority.

To decide who serves the community best we should also explore who does the most for suppressing the DC-type of behaviour, which is an obstacle to increasing societal wealth. In this context, the most effective means – aside from the demonstration effect and pressure exerted during negotiations – was the educational effect of the responses that reflect the partner's C or D, that is reactions that punished defection and rewarded co-operation. A closer look at the pair competitions reveals that the DC member of the 1 DC-FTFT pair turned into a TFT by the end, of 6 DC persons who were part of DC-TFT pairs

2 became DTFT and two TFT-type players, of the 6 who were in DC-DTFT couples 2 became DTFT and one turned into a TFT, while among the 4 DC only pairs one person became a TFT and two became DTFTs. Though the small numbers caution us from far-reaching conclusions, it is clear that TFT strategies "educate" players the most effectively.

All this supports the claim that in a DTFT (and DC) environment – both in the context of the experiment, and more generally, given the conditions of today's Hungary – TFT not only provides more success on the individual level, but also does most for communal well-being. A DTFT or CD attitude is more easily "educated" by the realistic TFTs warning defections (or rewarding co-operation), than by the FTFT's well-intentioned and forgiving („turn the other cheek") responses.[32] It seems therefore that the metaphorical "stick" is not only a necessary tool of the individual's road to happiness, but also an inescapable means of achieving the community's well-being. As another experiment revealed, the player who is willing to punish altruistically, that is she is willing to administer a punishment that will hurt her individually to penalize a transgression of the communal norms, is also a key asset to the community's well-being.[33]

In lieu of a conclusion

The experiment showed that the goal of creating an efficient yet humane community is better served by individuals who are committed to enforce their interests and in fact pay retribution for transgressions, than by individuals who benignly overlook others' parasitism and tolerate the violation of communal roles. TFT enhances community happiness more than either FTFT or CD. If we want to create a successful and wealthy community in Hungary today, we should neither follow Mother Theresa nor Gordon Gekko. When

[32] The "educational" process conducted by TFT is astonishingly reminiscent of the token economy method. Token economy is a behaviour-modifying method which reacts immediately to certain types of behaviour by punishment or reward, thereby showing whether the given behaviour was desired or whether it should be avoided. In their book *Mindwatching* Hans and Michael Eysenck describe the effect of immediate feedback between action and corresponding punishment or reward in the context of a token economy type prison education experiment. See Hans and Michael Eysenck. Elmevadászat. Kairosz Kiadó. 2002. Pp. 364-365.

[33] **Fehr cikk**

educating our children we would be better advised to pass on Hillel's advice: „If you are not for yourself, who is for you then? If you are only for yourself, who are you?" In my interpretation this means, that if you do not stand up for your interests, nobody else will. But you will not be any more successful by completely disregarding others' interests. As the TFT proved most efficient in influencing other players to respect communal norms, under the current conditions this type of behaviour does the most for creating a society that generates "wealth" and the opportunity to maximise "happiness points" for its individual members.