Evolution of Trust in IIIT-H

Dhruvee Birla, Raj Maheswari
2019115008, 2019101039
International Institute of Information Technology
Hyderabad, Telangana, India

Abstract—**Trust, a feeling that has more power over our decisions than we realize. It impacts the decisions we make in a societal setting, changes our decisions depending on who we are interacting with and therefore it is very important to understand the role it plays in a real world setting. Our aim is to acknowledge the presence of trust analogous to the 'Evolution of Trust' game but in a real world scenario. We play this game with the students of IIIT-H community to understand the (mis)trust epidemic, recognize the various strategies used by the participants while playing either with their friend or with a stranger, and discern the differences between the results of Nicky Case's interactive game and when the game is played among humans.**

I. Introduction

Evolution of Trust is an interactive game designed by Nicky Case [1]. It is based on the repeated Prisoner's Dilemma and has been highly inspired by Robert Axelrod's Evolution of Cooperation. The objective of our experiment is to observe and analyse human behaviour when we play this game among human participants. In the past, many iterated prisoner's dilemma tournaments have been played [3], where experts from disciplines of game theory, mathematics, economics, etc. have tried to come up with strategic programs to compete in a round-robin tournament. Through this project, we try to recreate Nicky Case's 'Evolution of Trust' game in a real world scenario. We observe how humans actually interact with each other when presented with this dilemma and use each others behavioral tendencies to form strategies.

Robert Axelrod's 'The Evolution of Cooperation' (Axelrod, 1984) [2] has been an inspiration for our project. We will be focusing not only on the aspect of cooperation but also on the aspect of trust. It is obvious that for maximum mutual gain, both participants would need to cooperate, but how much trust would players put in each other that their opponent does not cheat for their personal gains? Choosing to cheat has mainly 2 reasons. First, players try to take advantage of their opponents trust. This mentality only looks at short-term benefits and does not realise that the opponent will respond accordingly in the following rounds. The second reason to cheat, and a more motivating one, is the fear of getting cheated. In short, "Fear is a stronger motivator than greed." Greed does not produce long term benefits so humans may not be inclined to cheat often, however, once a person gets cheated on, they are less likely to trust their opponent again. Different players have different thresholds of tolerance of trust. Some are more forgiving and allow their opponents to take advantage of them for a few rounds, while others may be more suspicious and can hold grudges for a long time.

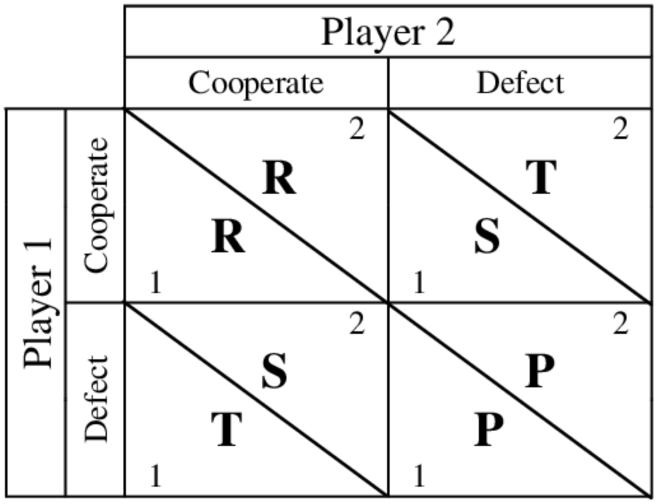


Fig. 1. Payoff-matrix with: R = 2, P = 0, T = 3, S = -1

A. Repeated Prisoner's Dilemma

The Prisoner's Dilemma is a thought experiment which places 2 rational agents into a dilemma. It is a commonly used concept in game theory where each player's best interest is to cheat their partner for maximum personal benefit. The dilemma is this: if they both cheat, the players are at a mutual loss compared to if they cooperate. However, cooperating carries with it a risk of getting betrayed. The payoff matrix as shown above, makes it clear why cheating is the most rational choice for both players. The Nash Equilibrium here, results in a sub-optimal outcome for each player.

However, in the repeated version of the game, there is a twist. Since the same scenario is presented to the player's multiple times, cheating in each round will NOT benefit in them in the long run. It may be advantageous to cheat when the game is played only once, but the goal here is to maximize their score across multiple rounds in interactions with many different players. As players play with each other, they will understand patterns in each other's behaviours and accordingly alter their choices.

A lot of research has been done in this field to understand the various different strategies possible when considering an infinite game. Since there are an unknown number of rounds, there exists no fixed optimum strategy. Prisoner's Dilemma tournaments have been widely played in the real world. The famous game 'Split or Steal' on TV show 'Golden Balls' is one such example where participants face this dilemma. In our game, we analyze participant's behavior and place them under one of the categories that have been researched till now.

II. Related Work

In ”Humans vs Bots”, S. Swarup et al., have studied human behaviour in the IPD (Iterated Prisonners Dilemma) by having participants play against software bots [5]. They observed that TFT (Tit-for-Tat) promotes cooperation while RTFT (Reversed-Tit-for-Tat) promotes defection. For each round, the subjects played against a randomly chosen bot. Our game differs from this in 2 ways. First, humans play against other humans, not bots. Although bots can also be programmed to play tactically and make errors sometimes, they can never completely replicate human behaviour. Moreover, the knowledge that your opponent is a human and not a machine is bound to change the way player perceive and make choices. Second, in our experiment, every player plays all rounds with only one opponent in a particular game. They do not play simultaneously with other players. This allows participants to remember their history of interactions and make informed choices. The ”Humans vs Bots” study concluded that either humans have not evolved behavioral strategies to deal with IPD situations or if they have evolved than the adaptation is reputation. This enables TFT strategy, and if it is hard to keep track of reputation, then MW (Majority Wins) is a fallback strategy.

According to Axelrod, Tit-for-Tat is successful because of its 2 properties- niceness and forgiveness [3]. It starts off with cooperating showing it is nice in the beginning, but also forgiving as it returns to cooperating once its opponent starts cooperating. It is not foolish to let its opponent take advantage of it. In our experiment, human emotions such as grudge and fear play an important role in the choices made. Also, since Tit-for-Tat is such a common strategy, players were inclined to use its variations in their games.

III. Experimental Design

The null hypothesis of this study is that ”participants will cooperate more with friends and cheat more with strangers.” Friends are likely to trust each other and have less fear of getting cheated. The alternate hypothesis of this project is that ”majority of the participants when put in a variant of repeated Prisoner's Dilemma will act either as a Detective or as a Grudger.”

Players know the identity of the person they interact with so they tend to play with different strategies depending on the degree to which they are acquainted with their opponent. Not every player gets a chance to interact with every other player.

The payoff matrix has a negative value when cooperate but your opponent cheats. Here we use the framing effect. This negative value invokes the bad feeling of getting betrayed.

A. About the Experiment

To create a real world scenario, we need 2 players playing with each other. They have two options - either to cheat or to cooperate. We note down the quantitative binary option they play as either 1 (cooperate) or 0 (cheat) and calculate the scores of each of them in the following manner :

- If both of them cheat, they get 0 points each
- If both of them cooperate, they get 2 points each
- If one cheats and one cooperates, the one who cheats gets 3 points and the one who cooperates loses 1 point

Both the players will be playing a terminal-based game which gives them two options - cheat or cooperate. Player1 and Player2 input the move they want to play simultaneously. After playing the first round, both the terminals show the move played by the other player. This way the players can now decide their moves based on the previous moves played by their opponent. Both the players only interact through the terminal and are kept in separate rooms to avoid other ways of communicating. Communication is prevented so players can not bargain or negotiate with each other. Players only have the history of choice to decide their next move. Although negotiating is usually practiced in real world scenarios, involving such additional factors can add noise to our analysis and is hence avoided.

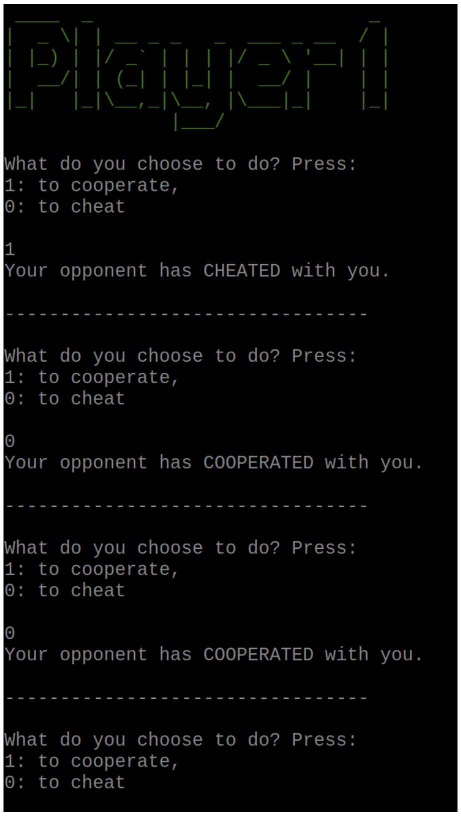
Players keep playing for some x number of rounds where what the ‘x’ is will not be known to the players but only to the researchers. This is done to simulate the infinite iteration of the game for the players. If players know in advance when the game will end, it is rational for them to betray each other according to backward-induction.

Each player will play 2 or more games. They will play at least 1 game with a stranger and 1 game with a friend so we can interpret if differences exists between their behaviors attributed to feelings for their partner. A player plays only one game at a time. In other words, while two players are playing together, no other game in parallel would involve these 2 players. This has been done to ensure focus to one game and for the player to efficiently decide on a strategy optimal for each game.

We conducted this experiment with 15 students, who belonged to the age group 17-22, where not all participants played more than 1 game. We played a total of 30 games. Using these 30 games we analyzed the strategies used by the individual players and placed them under the following 8 categories - CopyCat, Cheater, Cooperator, Grudger, Detective, CopyKitten, Simpleton and Random.

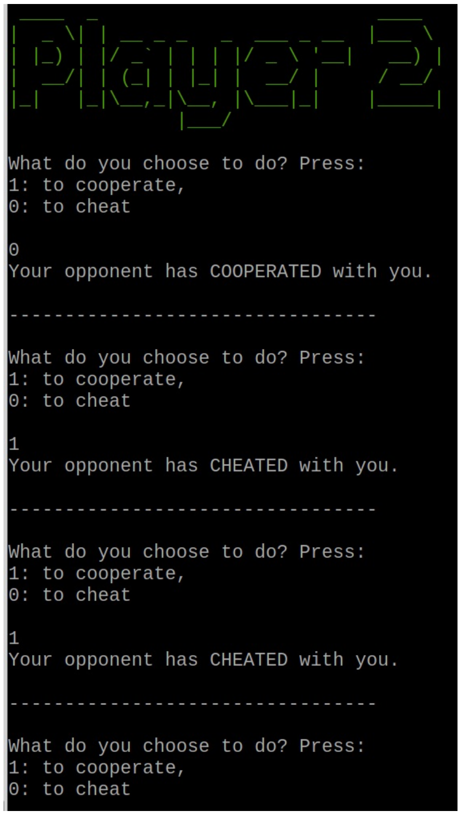
Their choices of play - cheat or cooperate throughout the games and how they change their strategies as the game progresses helped us analyze their human behaviour and relate them to the above mentioned categories.

Since there will be no incentive given to the player, they are told that the one who scores the most will win the game which we are hoping acts as a social incentive. The scores will be told after all the players finish playing at least 2 games. The reason why we are avoiding to reveal scores after every game is because the scores might have an affect or create a bias in the player's future games and we do not want them to form biases in real life based on the game play. We want the participants to play according to their inherent biases (which



```
 ____  _                       _ 
|  _ \| | __ _ _   _  ___ _ __/ |
| |_) | |/ _` | | | |/ _ \ '__| |
|  __/| | (_| | |_| |  __/ |  | |
|_|   |_|\__,_|\__, |\___|_|  |_|
               |___/             

What do you choose to do? Press:
1: to cooperate,
0: to cheat

1
Your opponent has CHEATED with you.

--------------------------------

What do you choose to do? Press:
1: to cooperate,
0: to cheat

0
Your opponent has COOPERATED with you.

--------------------------------

What do you choose to do? Press:
1: to cooperate,
0: to cheat

0
Your opponent has COOPERATED with you.

--------------------------------

What do you choose to do? Press:
1: to cooperate,
0: to cheat
```

Fig. 2. Terminal : Player1



```
 ____  _                         ____  
|  _ \| | __ _ _   _  ___ _ __  |___ \ 
| |_) | |/ _` | | | |/ _ \ '__|   __) |
|  __/| | (_| | |_| |  __/ |     / __/ 
|_|   |_|\__,_|\__, |\___|_|    |_____|
               |___/                   

What do you choose to do? Press:
1: to cooperate,
0: to cheat

0
Your opponent has COOPERATED with you.

--------------------------------

What do you choose to do? Press:
1: to cooperate,
0: to cheat

1
Your opponent has CHEATED with you.

--------------------------------

What do you choose to do? Press:
1: to cooperate,
0: to cheat

1
Your opponent has CHEATED with you.

--------------------------------

What do you choose to do? Press:
1: to cooperate,
0: to cheat
```

Fig. 3. Terminal : Player2

we believe will be different for a friend and a stranger) and do not want to add bias to the player's strategies for instance by revealing the scores after every game.

B. Infinite Iteration of the game

The game is played without a deterministic end, this is done to give a sense of infinite iterations. The infinite iteration of the game, simply put, means that neither of the players will know when they are playing their last round which is essential in a Iterated Prisoner's Dilemma.

By knowing the last round, both the players would know any move of theirs will not be consequential in the last round and they might both cheat. Similarly, knowing in the second last round that their move will not change the following round, they will again both cheat. This would happen for the third-last, fourth-last and so on for every round thus making this game unnecessary and inconsequential.

Hence, we have ensured that none of our players know what is the last round in their game and to keep it random, all pairs of players will have different number of rounds. However, players know that realistically the game is played somewhere between 7 to 15 rounds. Therefore, we believe that due to this, they play differently after the 9th or 10th round as if the game were to end.

C. Strategies

This sections mentions the different strategies that have been documented a player can use [4]. In our analysis and findings section we mention strategies used by participants which do not fall under any of these categories.

- **Cheater:** They cheat in all the rounds and all the games no matter the move of the other player.
- **Cooperator:** They cooperate in all the rounds and all the game no matter the move of the other player.
- **CopyCat:** In this strategy, the person starts with cooperating and then just copy whatever the other person did in the last round. In other words - Tit-for-Tat. Another variant of this is to start with cheating, called suspicious TFT.
- **Reverse Tit-for-Tat:** This is same as above, except it does the exact opposite of what the other person did in their previous round.
- **Grudger:** They start by cooperating and continue cooperating until you cheat them. Once you cheat them, they will always cheat. Another variant is that they will cheat for the next x rounds once they are betrayed.
- **Detective:** They start with analyzing the other player. Their first few moves is as follows - cooperate, cheat, cooperate, cooperate. If the other player cheats back, they will act like a copycat and if the other player never cheats back, they will always continue cheating. This strategy is also known as Prober.
- **CopyKitten:** They are like copycats and they cheat back only after the other player cheats them twice in a row thinking the first 'cheat' move could be a mistake.

- **Simpleton:** They start with cooperating. If the other player cooperates back, they play the same move as their previous move but if the other player cheats back, they play the opposite to their last move.
- **Firm But Fair:** They cooperate till the other person defects. Then, they resume cooperating after a round where both defect.
- **Soft Majority:** They cooperate as long as the number of times the opponent has cooperated is greater than or equal to the number of times it has defected, else it defects. Similarly, a hard majority is one which starts with defecting and defects in the equality condition.
- **Random:** These strategies are random, that is, there is a 50/50 chance to cooperate or cheat.

D. Stranger and Friend

As mentioned in the earlier sections, a player will play at least 2 games, one with a stranger and one with a friend. By a stranger we mean someone the player has not interacted with before and by a friend we mean someone who the player talks to on a day-to-day basis. Since IIIT-H is a small community we do not expect players to not know the other player at all, even if the players have seen each other before and are acquaintances, we consider them strangers.

Our reason for making players play with strangers and friends is to see whether their strategies change with changing players. This part of the project is not majorly focused on. Our primary aim still remains to understand the strategies used most commonly by the players which we expect to be a detective or a cheater. Our secondary analysis focuses on this aspect of the project, that is, whether the changing strategies are due to the relationships between the players or some other factors are at play.

IV. Analysis and Findings

The following analysis and our findings are based on the 30 games that our 15 participants played. This section is further divided into two sections - games played among friends (Friendly) and games played among strangers (Unfriendly). The total number of games played was 30 out of which 12 were 'Friendly' games and 18 were 'Unfriendly' games. Out of the games played, players were found equally likely to start with cooperating as cheating. This was seen to be true irrespective of whether or not they were friends.

A. Unfriendly Games

We notice a lot of participants initially start off playing with a random strategy but after playing a few rounds, they understand their opponent and come up with a scheme for their future rounds and games as well. For instance, one of the participants when playing with a stranger always cheated in their first game which is a naive strategy since it fails to account for the repeated nature of the game. The experimenters after the first game for all the participants explained the game to players once again specifying that the objective of the game is to maximize total score across all players. After this clarification, the participant who used the always cheat strategy earlier now starts cooperating in certain instances in another game with a stranger.

The CopyCat strategy was noticed to a great extent in participants playing with strangers. In one of the games, the players adopt the 'Suspicious Tit for Tat' and 'Tit for two Tats' strategies. Suspicious Tit for Tat is in essence the CopyCat strategy where the only difference is the player in this strategy starts by cheating instead of cooperating. Tit for two Tats is another term for CopyKitten where the player gives their opponent the benefit of the doubt assuming the first 'cheat' move was a mistake. In the latter game, the 'Grudger' strategy was noticed to be in use by both the players. In two other games, the players use the CopyCat strategy with 20% and 25% chances of deferring. The chances of deferring arise due to the formulation of a good strategy during the earlier rounds of the game. Once the strategy has been arrived at, participants continued using the CopyCat strategy till the end of all the rounds in the game. Another reason for Deferring is to probe the over-forgiving nature of the opponent.

We observed that some players tried cooperating quite often for the overall good despite being exploited by their opponents. Most players tried to predict the choice made by their opponent by drawing patterns from previous interactions. If they chose the option which the believed their opponent would take. In effect, their attempt was to have either both cooperate or both cheat. If one cooperated and other cheated, it would create distrust and be disadvantageous in the next rounds. The most common reasons for such inconsistencies was failure to predict the other persons choice because of lack in communication. Friends were better able to predict their partners moves than strangers.

The 'Grudger' behavior was very evidently noticed in one of the games where both the players played for 10 rounds and one of the participant consistently cooperated in the first 5 rounds and on the 5th round the opponent had cheated which led to the player consistently cheating till the end of the game. For the first few games, we noticed strategies like the CopyCat and other derivations of the CopyCat strategy. But we also noticed strategies which contradicted our null hypothesis. We noticed a game among strangers where two of the players always cooperated. Similarly, in the game among friends, one of the friends always cheated.

B. Friendly Games

Another interesting observation was made from a game played between friends. It was noticed that the two of them cooperated in 60% of their rounds while deceiving each other 15% of the time to take advantage of the other person cooperating. Both friends attempt to get extra points occasionally in lieu of their partner's trust. However, this is unlike games played between strangers where the players cheat excessively to avoid getting betrayed.

In some of the games, we noticed no matter how many times one player cooperates, the other player always continued cheating. They used the 'cheater' strategy against their friends

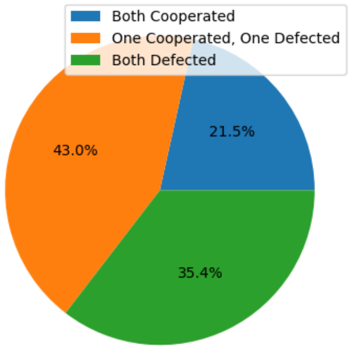


Fig. 4. Pie-chart showing fractions of possible outcomes from 223 total rounds.

to maximize their score. In another game, a player consistently cooperated for some moves and then consistently cheated in the remaining moves without any regard for the opponent's moves.

Around 10 games out of the 12 friendly games that were played, started with at least one of the players cheating. Implying either they both cheated in the first round or at least one of them cheated in the first round of the game. Additionally, there were also cases as mentioned earlier where one of the players continued cheating using the cheater strategy which disproves our null hypothesis.

V. Result

From the 30 games that were played as part of this experiment, we can draw 2 conclusions :

- There is no difference in the way players play with their friends and with strangers.
  - This result disproves our null hypothesis.
  - We have noticed that the main goal for the player becomes maximizing their scores across all players. They do not cooperate more with friends or cheat more with strangers. They play according to what strategy suits best at the time of the game with that particular opponent.
- The most common strategies adopted are CopyCat, Grudger and Random.
  - This hypothesis is consistent with our alternate hypothesis.
  - A few derivations of the CopyCat strategies are also noticed namely CopyKitten aka Tit for Two Tats and Suspicious Tit for Tat where the players starts off with cheating and then plays as a CopyCat.

VI. Conclusion

The main aim of our study was to recreate Nicky Case's Evolution of Trust game in a real world scenario. To do the same, we created a simpler version of the Evolution of Trust game in the form of a terminal-based game where the players had to input either 1 or 0 meaning cooperate and cheat respectively. The terminals displayed the opponent's move on the screen for the player to make decision for their next round. The players were told who they will be playing against since we wanted to also analyze the difference between games played among friends and games played among strangers. However, the players were not told their scores or their position in the leader board so as to reduce biases in real life based on the game play. In the total of 30 games played, 12 were games played among friends and 18 were played among strangers. Our null hypothesis of noticing differences among Friendly and Unfriendly games was contradicted from the results obtained and have been explained in the earlier section. But we do notice the common strategies to be that of CopyCat, Grudger and Random which we had hypothesized.

VII. Limitations and Future Work

In Nicky Case's Evolution of Trust game, we had found that CopyCat or Tit-for-Tat as a strategy was one of the most efficient and common strategies and so can be seen as a strategy used consciously or subconsciously by many of our participants in this study. We believe that some participants may already have been aware of this game and thus would have known the best strategy to maximize their score. Furthermore, we believe that the random strategy would have been used much less often if we would have provided them with monetary rewards instead of a social incentive. Even though the social incentive is expected to produce as much performance enhancement as monetary reward, we believe so was not the case in our study since the participants felt that the leaderboard in general provided them with no rewards. As a basic law of human behavior we know that the more the incentives, the more is the effort and performance [6]. Therefore, we are of the opinion that if the incentives would have been more then getting a higher score would encourage performance enhancement, resulting in obtaining more coherent results with less use of Random strategy.

We also noticed that participants of the same year discussed their strategies and ways to optimize their moves. Although such discussion has no direct effect as players are free to do whatever they want, it should have been avoided so that players have a fresh perspective in each game played by different players.

Due to time constraint and the scope of this project, we were not able to get more than 15 participants. As a future development of this project, we suggest the following tasks:

- Make some variations in the game so that the participants are not already familiar with the game and get a chance to explore the best strategies during play.
- Robert Axelrod's Evolution of Cooperation (Axelrod, 1984) can be studied in depth and inferences can be drawn from the cooperation aspect as well. These inferences can then be used to perform more comparisons

between trust as a trait and cooperation as a trait since in our study we did not consider trust to be a trait.

- If trust is considered as a trait, one can perform questionnaires to form more in depth analysis on the role of trust.
- This game can be extrapolated to understand the general role of trust in human behavior and how it is different from other aspects of human behavior. In addition to this, one can also notice the role of trust in general human decision making.
- Keep the participants in separate rooms to avoid communicating their ideas.
- Provide the participants with monetary incentives.
- We can also experiment with the participants using altered payoff values.

Due to time constraint and the lack of resources, we could not complete the above mentioned tasks but we believe those tasks can help us perceive ’trust’ as a trait and understand how significant its role is in human behavior.

References

[1] https://ncase.me/trust/

[2] Axelrod, R., Hamilton, W. D. (1981). The evolution of cooperation. Science, 211(4489), 1390–1396.

[3] https://egtheory.wordpress.com/2015/03/02/ipd/

[4] https://medium.com/thinking-is-hard/a-prisoners-dilemma-cheat-sheet-4d85fe289d87

[5] S. Swarup, M. G. Orr, G. Korkmaz and K. Lakkaraju, ”Humans vs. Bots: investigating models of behavior in the iterated Prisoner’s dilemma,” 2020 Spring Simulation Conference (SpringSim), 2020, pp. 1-12, doi: 10.22360/SpringSim.2020.HSAA.009.

[6] https://nectarhr.com/blog/the-difference-between-monetary-non-monetary-rewards