The N-Player Trust Game and its Replicator Dynamics

Hussein Abbass, Garrison Greenwood, and Eleni Petraki

Abstract—Trust is a fundamental concept that underpins the coherence and resilience of social systems and shapes human behavior. Despite the importance of trust as a social and psychological concept, the concept has not gained much attention from evolutionary game theorists. In this paper, an N-player trust-based social dilemma game is introduced. While the theory shows that a society with no untrustworthy individuals would yield maximum wealth to both the society as a whole and the individuals in the long run, evolutionary dynamics show this ideal situation is reached only in a special case when the initial population contains no untrustworthy individuals. When the initial population consists of even the slightest number of untrustworthy individuals, the society converges to zero trusters, with many untrustworthy individuals. The promotion of trust is an uneasy task, despite the fact that a combination of trusters and trustworthy trustees is the most rational and optimal social state. This paper presents the game and results of replicator dynamics in a hope that researchers in evolutionary games see opportunities in filling this critical gap in the literature.

Index Terms—Trust, Evolutionary Game Theory, Trust Game, N-Person Trust Game

I. INTRODUCTION

Trust is the glue of a social system [24]. Despite its vital role in the society, the concept is absent from the evolutionary computation literature. When compared to the large number of papers published on the iterated prisoner dilemma [13], [6], [21], [2], [27], [23], there has not been any publication on trust. The contributions of this paper are three-fold. First, it aims to encourage more work on trust in the evolutionary computation and evolutionary game theory research areas. Second, it introduces a novel N-player trust game that assists researchers in understanding and analyzing the concept of trust using evolutionary game theory. The game is a social dilemma and generalizes the concept of trust, which is normally modelled as a sequential game, to a population of players that can play the game concurrently. Third, it presents the first theoretical analysis of the dynamics of the game using replicator dynamics.

A recent review on trust in social and psychological literature [30] demonstrated that there are abundant studies on the roles of trust [10], [11], [24] and its implications for social and human systems [12], [14], [17], [31], including ethical considerations [3] surrounding studies of trust and the means for influencing and shaping trust [5], [26]. Managers perceive how an understanding of how trust is formed can create opportunities to develop loyal customers [4] and improve relationships among employees and management [19], [33], [32]. Individuals perceive trust as a vulnerability [26] in which one person exposes himself to another by relying on the other person to make a decision on his/her behalf, as an opportunity [28] to create favorable outcomes, and as a source of unwanted uncertainty [16] that creates a relationship in which the trustee has power over the trusting party.

In the social and psychological literature, trust is described as playing two important roles. First, sociologically, it acts as a complexity-reduction mechanism [24], allowing individuals with limited cognitive capacities to manage the complex world they live in. Luhmann [24] showed that trust creates a positive feedback loop in a social system. As individuals use trust to manage complexity, relationships emerge, a process that increases complexity and thus creates more reinforcement and opportunities for trust to spread.

Second, psychologically, it acts as an ambiguity- and uncertainty-reduction mechanism for individuals [9], [11]. Deutsch [11] argues that a trusting situation occurs when a truster perceives that a situation has one of two potential outcomes, where one is perceived to have a negative valency of greater absolute value than the positive valency attributed to the second. However, which outcome will occur depends on the trustee. If the truster chooses to proceed, the truster is said to trust the trustee; otherwise, the truster distrusts the trustee. Context influences the perceived levels of both trustworthiness and risk; as the risk of trusting changes from one context to another, the decision to trust changes as well. Therefore, trust and risk are tightly coupled concepts.

Trust games are sequential in nature. The truster must first decide whether to trust the trustee. If the decision is to trust the trustee, the trustee must then decide whether to be trustworthy or not. Each decision can be a binary decision or be considered as part of a continuum representing the degree of either trust or trustworthiness. Although the proposed game relies on binary decision-making, it can easily be generalized to include non-binary decisions. Neuroeconomic experiments use two main forms of trust games [1], [9], [16]. We will call them TG1 and TG2 for trust game 1 and 2, respectively.

TG1 [22] is a non-zero-sum game, and the trustee must choose whether to trust the trustee. If the trustee chooses not to trust the trustee, both players receive $5. If the trustee chooses to trust the trustee, then the trustee must decide whether to be trustworthy or not. If the trustee chooses to be trustworthy, the truster and trustee receive $10 and $15, respectively. If the trustee chooses to defect, the truster gets nothing, whereas the trustee gets $25.
In this game, a rational truster would be indifferent to the two choices under the assumption that the trustee’s probability of being trustworthy is 0.5. Therefore, the expected return for the truster, regardless of the decision to trust or not, is $5. The trustee, however, benefits more if the truster chooses to trust the trustee. In this case, in a one-off situation, a rational trustee would choose to defect to maximize his or her own return. However, over repeated iterations, the trustee has an incentive to be trustworthy because being so will yield maximum returns to both the trustee and truster. The game tree for this game can be seen in Figure 1.

TG1 can be generalized, we call it TG1G, to the game tree shown in Figure 2. An early version of this general form was presented in [25] and the general version was then analyzed in [26].

TG2 [1] is also a non-zero-sum game. Both the truster and trustee start with an endowment of $12. The truster makes the first move by deciding how much money to transfer to the trustee. The money is tripled on the way to the trustee. The trustee then must decide how much money to transfer back to the truster. If X is the amount of money the trustee sends back to the truster and Y is the amount of money the truster sends back to the trustee, then the truster will end up with a balance of $12 - X + Y while the trustee ends up with a balance of $12 + 3X - Y. Both the truster and trustee end up with the same balance when Y = 2X. In this game, X is a measure of trust and Y is a measure of trustworthiness. This game is represented in the following table.

Neither TG1 nor TG2 create a complete social dilemma. If the truster chooses to trust the trustee, then the total wealth, independent of whether the trustee is trustworthy or not, is $25 in TG1 and 24 + 2X in TG2. The social dilemma, therefore, is reliant only on the truster’s decision. However, in TG1G, the game is a social dilemma when 0 < r < 1.

## II. N-PLAYER TRUST GAME

Assume N players. Each player must make two decisions in advance. The first is to decide whether to be trustworthy or not. The second is to decide whether to govern (be a governor or a trustee) or be governed (be a citizen or a truster). We will denote the former as G and the latter as C. Assume the number of players that decided to be type C is \( x_1 \) and the number of players that decided to be type G is \( x_2 + x_3 \); such that, \( N = x_1 + x_2 + x_3 \).

In real-world settings, the role of the truster and trustee are defined *a priori*. For example, an investor chooses to invest, while a financial planning advisor chooses to become one before the two actors decide to enter into a trusting relationship. This type of decision we refer to as a social choice. The second type of decision normally decided *a priori* as well is whether the trustee will be trustworthy or not. The reason it is ‘normally’ decided *a priori* is that it largely depends on the behavioral attitude and value system of the trustee. We acknowledge that context influences this behavioral attribute but we conjecture that this decision is fundamentally a core behavioral attribute of the agent. We call this decision a behavioral choice.

In summary, it is reasonable to assume that the two types of decisions related to social and behavioral choices can be done *a priori*.

A player of type \( C \) pays \( tv \) to the government, where \( tv \) denotes the trusted value. The dynamics of the game is in general independent of the value of \( tv \), which can be set to 1. However, we maintain \( tv \) to allow flexibility in adopting the game to different contexts. With \( x_1 \) players of type C, the total money that is sent to the government is \( x_1 \cdot tv \). Each player of type G receives \( (x_1 \cdot tv)/(x_2 + x_3) \). Assume \( x_2 \) players of type G decide to be trustworthy, while \( x_3 \) decide to not be trustworthy. A player in the \( x_2 \) population returns to the citizens a multiplier of \( R_1 \) of what was received and keeps the same amount for himself, with \( R_1 > 1 \). A player in the \( x_3 \) population returns nothing to the citizens and keeps for himself a multiplier of \( R_2 \) of what was received, where \( R_1 < R_2 < 1 \).

<table>
<thead>
<tr>
<th>Starting Balance</th>
<th>Truster</th>
<th>Trustee</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pay</td>
<td>X</td>
<td>Y</td>
</tr>
<tr>
<td>Receive</td>
<td>Y</td>
<td>3X</td>
</tr>
<tr>
<td>Net Wealth</td>
<td>12 − X + Y</td>
<td>12 + 3X − Y</td>
</tr>
</tbody>
</table>

### TABLE I

**Utility Matrix for a 2-Player Trust Game.**
Recall that $R_2 > (2R_1 - R_2) > 0$, $N$ is a constant, and $x_2 + x_3 \geq 1$, the maximum of the function will occur when $x_2 + x_3$ is at a minimum ($x_2 + x_3$ appears twice, once with a negative sign in the numerator and once in the denominator) while $x_3 = 0$ is the maximum for $\frac{x_2}{x_2 + x_3}$; therefore, $x_2 + x_3 = 1 \Rightarrow x_2 = 1$.

**Theorem 3:** $x_1 = N - 1$ and $x_2 = 1$ is Pareto Optimal.

**Proof.** The single $x_2$ player can switch to an $x_3$ player or an $x_1$ player; either change reduces the $x_1$ player payoff to zero. All investments are split among the $x_2$ and $x_3$ players so any $x_1$ player that changes to an $x_2$ or $x_3$ player reduces the payoff to the single $x_2$ player. The proof follows that any player unilaterally switching roles reduces the payoff to an $x_1$ player.

Our investigations indicate other Pareto Optimal solutions exist, but the one in Theorem 3 is known as a social welfare maximizer or socially optimal solution because it maximizes the total utility for the population.

If every player chooses to be a governor and untrustworthy, the net individual wealth is zero. In contrast, if a player chooses to be in the $x_1$ population, he or she can lose all of their money if all of the governors are untrustworthy. Would a single trustworthy governor emerge in this population to maximize the combined wealth of the population?

The evolutionary behavior of a population playing the trust game can be studied using replicator dynamics \[20\]. Let $y_k$ be the frequency of the $x_k$ players in an infinitely large population with $\sum_k y_k = 1$. Then the time evolution of $y_k$ is given by the differential equation

$$\dot{y}_k = y_k \cdot \left( f_k - \bar{f} \right)$$

where $f_k$ is the expected fitness of an individual playing strategy $k$ at time $t$ and $\bar{f}$ is the mean population fitness. Here, fitness and net wealth are equivalent. The number of copies of a strategy increases if $f_k > \bar{f}$ and decreases if $f_k < \bar{f}$. We can calculate $\bar{f}$ as follows

$$\bar{f} = \frac{y_1 \cdot y_2 \cdot tv \cdot (2 \cdot R_1 - 1) + y_1 \cdot y_3 \cdot tv \cdot (R_2 - 1)}{(y_2 + y_3)}$$

The three replicator equations (Note that $y_2 + y_3 = 1 - y_1$)

$$\dot{y}_1 = \frac{y_1^2}{1 - y_1} \cdot \left( y_2 (1 - 2 \cdot R_1) + y_3 (1 - R_2) \right) + \frac{y_1 \cdot tv}{1 - y_1} \cdot y_2 (R_1 - 1) - y_3$$

$$\dot{y}_2 = \frac{y_1 \cdot y_2 \cdot tv}{1 - y_1} \cdot \left( y_2 (1 - 2 \cdot R_1) + y_3 (1 - R_2) + R_1 \right)$$

$$\dot{y}_3 = \frac{y_1 \cdot y_3 \cdot tv}{1 - y_1} \cdot \left( y_2 (1 - 2 \cdot R_1) + y_3 (1 - R_2) + R_2 \right)$$

2R1. The payoff matrix for this game can then be represented as shown in Table II with the following constraints:

$$1 < R_1 < R_2 < 2R1$$

$$N = x_1 + x_2 + x_3$$

The game tree is presented in Figure 3 for an arbitrary number of players.

**Theorem 1:** $x_3 = N$ is a Nash Equilibrium.

**Proof.** The order of the net individual wealth in the table from left to right is a monotonically increasing sequence. A player deciding to play according to any strategy on the right-hand side will yield a better net individual wealth than any of the strategies on the left-hand side. The best net individual wealth occurs for players in $x_3$. Any rational player will choose to be a governor and untrustworthy. Thus, the Nash equilibrium occurs when all players choose to be governors and untrustworthy.

The total combined wealth $CW$ of the population is

$$CW = \begin{cases} x_1 \cdot tv \left( \frac{2R_1 - x_2}{x_2 + x_3} + \frac{R_2 - x_3}{x_2 + x_3} - 1 \right) & \text{if } x_2 + x_3 \geq 1 \\ 0 & \text{otherwise} \end{cases}$$

(3)

**Theorem 2:** The maximum of $CW$ occurs when $x_2 = 1$ and $x_3 = 0$.

**Proof.** Given the constraints on $R_1$ and $R_2$, $CW$ has a lower bound of zero regardless of the value of $x_2 + x_3$. Therefore, to maximize $CW$, we only need to focus on the case when $x_2 + x_3 \geq 1$. Substituting $x_3 = (x_2 + x_3) - x_2$ into Eq. (3),

$$CW = x_1 \cdot tv \cdot \left( \frac{2R_1 - x_2}{x_2 + x_3} - 1 + R_2 - \frac{R_2 x_2}{x_2 + x_3} \right)$$

$$= x_1 \cdot tv \cdot \left( \frac{(2R_1 - R_2) x_2}{x_2 + x_3} + (R_2 - 1) \right)$$

$$= (N - (x_2 + x_3)) \cdot tv \cdot \left( \frac{(2R_1 - R_2) x_2}{x_2 + x_3} + (R_2 - 1) \right)$$

Fig. 3. Game tree for the proposed N-player trust game.
Figure 4 shows the population evolution for various initial player distributions. The replicator equations predict there is a critical ratio of $x_1$ to $x_2$ players that acts as an attractor where players rapidly switch strategy to become untrustworthy. However, untrustworthy players do not completely take over the population because a certain proportion always remains trustworthy even when there are a few or no investors left. Without investors—i.e., $y_2 = 0$—each replicator equation has the form $\dot{y}_k = 0$, which is a fixed point. Importantly, when there are no longer any investors, the net worth of all remaining players is zero. Consequently, there is no incentive to switch strategies and a steady-state condition exists.

The bottom trajectory in Figure 4 shows an extreme case that is highly favorable to a small number of trustworthy players: a population composed almost entirely of investors with almost no untrustworthy players. Under these circumstances, the net worth of a trustworthy player is much higher than that of an investor; therefore evolution should favor the trustworthy players. Untrustworthy players have an even higher net worth, but they are initially rare in the population. The replicator equations initially predict very rapid growth in trustworthy players with a corresponding plummet in investors. However, the inevitable increase of untrustworthy players reverses the trustworthy player growth. Even in this extreme case, the steady-state condition is reached.

The replicator equations predict interesting behavior in the trust game. Regions to the left of the attractor have low $C$ to $G$ ratios. Since investments are split among $G$ players higher returns go to $x_1$ players if most $G$ players are trustworthy. The replicator equations predict $x_2$ players will mostly switch to $x_1$ players. Conversely to the right of the attractor there is a high $C$ to $G$ ratio. Under those circumstances there is a strong temptation to be untrustworthy and the replicator equations predict a sharper rise in $x_3$ players. Near the attractor the returns to $x_1$ and $x_2$ players are roughly the same so there is little incentive to switch from $x_2$ to $x_1$ or vice versa. The maximum return is obtained by becoming untrustworthy and the replicator equations predict virtually all strategy changes are to $x_3$ players. Eventually the trajectories intersect the $y_2 - y_3$ line where there are no investors. At that point all replicator equations are of the form $\dot{y}_k = 0$ and a fixed point is reached. This makes perfect sense because, without any investors, there is no incentive to switch strategies.

Figure 5 shows the effect of different $R_1$ and $R_2$ values (but with $R_1 < R_2$). A fixed point is still always reached, but the lower the values, the more trustworthy players in the final population. The reason is lower $R_1$ and $R_2$ values cause the $y_1$ players to go extinct quicker, which limits the growth of untrustworthy players. Notice the similarity of these trajectories to those in Figure 4, which suggests the presence of an attractor. However, although all attractors are oriented in the same direction, their location in the 2-simplex depends on the $R_1$ and $R_2$ values.

**Theorem 4:** $y_3 = 0$, $y_1 = \frac{R_1 - 1}{2R_1 - 1}$ and $y_2 = \frac{R_1}{2R_1 - 1}$ is a
forms of two-player games to model trust and described the relationships between trusting decisions and human neural functions. These games do not create a social dilemma where the total social wealth depends on the decisions of both truster and trustee, and they do not generalize to multiple players. The field of evolutionary computation has seen little or no work on trust, despite the socioeconomic and psychological significances of the concept.

In this work, we introduce a new N-player trust game that can model trust decisions among many players. It creates a social dilemma in which individuals who attempt to maximize their own benefits in the short run maximize neither the society’s social wealth nor their own benefits in the long run. The results revealed that, while the optimal solution for the population exists when N-1 players choose to be trusters and the remaining players choose to be trustees, this optimal solution is a needle in a haystack, causing the evolutionary dynamics to consistently converge to a population without any trusters and with a combination of trustworthy and untrustworthy individuals. The exception to this phenomenon occurs when the initial population is free of untrustworthy players. The proposed game shows that trust is an uneasy but worthwhile concept.

**ACKNOWLEDGEMENT**

Portions of this work was funded by the Australian Research Council Discovery Grant number DP140102590.

**REFERENCES**


