CHAPTER 14

EVOLUTION AND BREAKDOWN OF TRUST IN CONTINUOUS TIME

AMNON RAPOPORT AND RYAN O. MURPHY

Scholars have long studied trust and what creates trusting behavior, generally with an aim to increasing trust and capturing the efficiency gains that are likely to result. How does trust come about, how is it sustained, and when does it break down? Amnon Rapoport and Ryan O. Murphy address these questions in the context of a general trust game that evolves over time. The game they present has the potential to explore elements of trust not captured by the games that researchers have used in the past.

Amnon Rapoport is Distinguished Professor of Management at the University of California Riverside's School of Business Administration. He is one of the pioneers of experimental and behavioral methods in decision science. Ryan O. Murphy is Professor and Chairperson of decision theory and behavioral game theory at ETH Zürich. His research focuses on human decision making in both individual and strategic contexts.

* * *

Research paradigms for studying trust and trustworthiness

Negotiations involve multiple individuals who cooperate to arrive at a joint decision that entails joint consequences, or payoffs, for each of them. Negotiations are the key to resolving conflicts. Negotiators make their moves over time in a relatively unstructured environment to identify the problem; to clarify their own objectives that include the hopes, needs, desires, and fears that motivate them; to generate creative alternatives that satisfy their preferences; to evaluate the consequences of the alternatives they generate; and to make tradeoffs (Raiffa, 2002). Trust is fundamental to this process of creating value, making tradeoffs, and rising above the confines of narrow self-interest in social settings.

Psychologists (e.g., Deutsch, 1960) and economists have proposed alternative paradigms to study trust and trustworthiness in the controlled environment of the laboratory. One of the most predominant paradigms in economics was established with the pioneering works of Dasgupta (1988); Güth and Kliemt (1994); and Berg, Dickhaut, and McCabe (1995). We start this chapter by describing this paradigm and critically discussing its major features. In particular, we note that this paradigm does not consider such design features as reputation building, incremental moves, changes in values, and signaling, all of which may foster and sustain mutually beneficial behavior over time. We then propose another paradigm, which we view as complementary, in which cooperative behavior is maintained by mutual trust that evolves continuously over time. This trust is fragile, as each individual may unilaterally terminate the interaction at any time and simultaneously increase her own payoff and decrease the payoff of each of the other individuals. After exemplifying and discussing this paradigm, the chapter summarizes the results of two preliminary experiments that focus on the effects of the incentive to defect, group size, and signaling on the evolvement of trust and proposes extensions of the basic model.

An experimental paradigm for studying trust and trustworthiness, which has dominated much of the research on trust in economics, originated with the works of Dasgupta (1988), Güth and Kliemt (1994), and Berg et al. (1995). What has become known as the investment game or the trust game (Berg et al.) has the following structure. There are two players, player i (called sender) and player j (called receiver). The sender is given an endowment (e.g., y_i = \$10). At the first stage of the game, the sender may transfer any amount of her endowment s_i ($0 \leq s_i \leq y_i$) to the receiver. Each dollar sent is exogenously tripled by the experimenter, so that the receiver—the second mover in the game—receives $3 \cdot s_i$. At the second and last stage of the game, the receiver can choose to return any amount r_j ($0 \leq r_j \leq 3 \cdot s_i$) to the sender. This concludes the game. All of these features are commonly known to both players. This very simple, two-stage extensive-form game is solved by backward induction. The receiver has no incentive to send any money to the sender

because returning money—a strictly dominated strategy for the receiver—necessarily subtracts from her own payoff. Anticipating no return payment from a rational receiver, the sender similarly has no incentive to transfer money to the receiver in the first place. For rational players, intent on maximizing their individual payoffs, the implication of the subgame perfect equilibrium is that the sender transfers nothing and, even if presented with the opportunity to respond, the receiver returns nothing. The subgame perfect equilibrium solution is Pareto deficient as both players could have earned more from the interaction had the sender sent her whole endowment and the receiver returned half of the tripled amount; however, maximization of social welfare would require both trust and trustworthiness from both players.

First-mover transfers of money have traditionally been interpreted as manifestations of trust, and second-mover return transfers as manifestations of trustworthiness. Numerous experimental and field studies of the trust game and many of its variants have provided evidence that solidly rejects the equilibrium prediction. Thus, for example, Berg et al. report that only five out of their 60 first movers transferred no money to the receiver. These findings have resulted in the conclusion that many, although not all, decision makers do not follow narrow self-interest dominant pure strategies, nor do they expect such behavior from their counterparts. Principles of backward induction, which play a fundamental role in the analysis of finitely iterated noncooperative games, fail to account for what seems to be trusting and trustworthy behavior. For a representative sample of experiments and field studies on trust and trustworthiness that implement different variants of the trust game, see Bacharach, Guerra, and Zizzo (2007); Burnham, McCabe, and Smith (2000); Cox (2002), Engle-Warnick and Slonim (2001), Glaeser et al. (2000); Güth, Ockenfels, and Wendel (1993, 1997); Ho and Weigelt (2005); McCabe, Rigdon, and Smith (2002); McCabe, Rassenti, and Smith (1996, 1998); McCabe, Smith, and LePore (2000); and Ortmann, Fitzgerald, and Boeing (2000). Camerer (2003) provides a comprehensive review of the experimental literature and Cardenas and Carpenter (2008) present an overview of field studies.

Murphy, Rapoport, and Parco (2006) have noted that the trust game creates an experimental setting with no institutional mechanisms conducive to fostering trusting or trustworthy behavior. The game provides no scope for personal relations or social networks but rather is typically conducted under conditions of complete anonymity. Perhaps more importantly, it eliminates other design features (e.g., repeated games that allow for reputation building, incremental moves, promises, threats, and the signaling of trustworthiness) that could foster and sustain mutually beneficial behavior over time (e.g., Dasgupta, 1988; Hardin, 2004; Kurzban, Rigdon, and Wilson, 2008). Other features of the trust game that may have reinforced an overly narrow definition of the concept of trust have also been noted. First, with few exceptions, studies of trust have focused on dyadic interactions. However, there is nothing in the various explications of the notion of trust (Rousseau et al., 1998; Fukuyama, 1995) that restricts it to two-player interactions. Members of economic alliances or scientific research teams formed to solve a

particular problem have to trust one another to complete their share of the joint project even without perfect monitoring. In these cases, mutual trust (Hardin, 2004) may be manifested in groups or teams with more than two members.

A second and possibly more critical limitation of the trust game is the built-in asymmetry between the two players. The asymmetry can prevent the receiver from registering any move at all. For example, in the subgame perfect equilibrium solution to the original trust game, the game terminates with the sender choosing to send nothing, therefore providing no opportunity for the receiver to register any decision, as she has no money to allocate. Among other implications, this feature of the game creates a demand characteristic, the effect of which may not easily be ascertained. Being recruited for and instructed to participate in an interactive decision-making task, the sender may discount the possibility of making a choice ("send nothing") that would essentially prevent the receiver from taking any action at all. Consequently, he may believe that it is expected of him to transfer at least some money so that the other player is provided the possibility of taking some action (otherwise, why would there be another player in the first place?).

Ideally, one would like to study the evolution and breakdown of trust in social interactions that are minimally encumbered by exogenously defined roles and demand characteristics. Hardin has commented that calling this paradigm a trust game is misleading if the game is not iterated in time (2004, 16). Rather, "[t]he prototypical case of mutual trust at the individual level involves an interaction that is part of a long sequence of exchanges between the same parties" (2004, 17). Dasgupta has similarly concluded that trust is based on reputation and that reputation has to be acquired through behavior over time (1988, 53).

Finally, there is the ambiguity in regard to the definition of what exactly the sender in the trust game is trusting. Is he trusting that the receiver will return at least the amount that he transferred to her? Is he trusting that the receiver will send back a prespecified proportion of the surplus created by the act of trusting and possibly based on some focal point, social norms, or some convention (e.g., 50 percent of the total amount after it is tripled, or perhaps the amount sent first plus 50 percent of the surplus)? Or less precisely, is he trusting that the receiver will treat him fairly? Another concern (Butler, Giuliano, and Guiso, 2010) is that, because the receiver neither makes any promise of what she would do nor has any agreement with the sender about how the two of them should divide the windfall, there is no scope for the sender to feel cheated by the receiver's action; hence, this leaves no role for trust in the trust game.

These critical comments are not intended to detract from the findings uncovered in dozens of studies of the trust game. Clearly, one-way trust relationships in dyadic interactions are of great analytical interest to researchers because of their simplicity and the tractability of their solutions. Also evidently, there are social interactions (e.g., parent-child, surgeon-patient) that are virtually one-way relationships and other contexts in which two parties have asymmetric roles (e.g., commander-soldier, manager-subordinate). However, we side here with Hardin who claims that the more stable, compelling, and interesting trust relationships

are likely to be mutual and ongoing. Player X_1 trusts players X_2, X_3, . . . , X_n because it is in the collective interest of each to do what she trusts them to do, and each trusts player X_1 for the same reciprocal reason. We do not consider the model that we present here as the definitive explication of the notion of trust nor do we claim that our model is the only one that examines aspects of trust that evolve over time. For example, there is a large body of literature on reputation building that is closely related to the experimental work summarized in part 3 (see, e.g., the chapter in this volume on negotiating reputations by Ockenfels and Resnick, and the chapter by Resnick and Zeckhauser, 2002, on trust in Internet transactions). We view trust as a construct akin to the concepts of knowledge, power, or belief and posit that any attempt to define trust exhaustively will unnecessarily restrict its scope and ultimately may cause more harm than good. Rather, our purpose is more modest, namely, to present and illustrate a relatively new paradigm—one of many possible—for experimentally studying the evolution, dynamics, and breakdown of mutual trust among a finite number of players.

The real-time trust game

Following Murphy et al. (2006), we focus on noncooperative games that evolve continuously over time in which there are *n* symmetric players, cooperative behavior is maintained by mutual trust, the joint payoff of trust-based cooperation grows over time, and any player may unilaterally[1] terminate the interaction at any time thereby enhancing her own payoff but subsequently decreasing the payoffs of each of the other players. To motivate this class of games and the kind of mutual trust that they are designed to model, consider a hypothetical vignette of *n* researchers who work on a problem of shared interest and decide to combine their interests, resources, and facilities to enter into scientific collaboration. This scenario is rather common in medical and biotechnological research (e.g., the Alzheimer's Disease Neuroimaging Initiative). None of the researchers is particularly familiar with the other team members, as all have worked mostly alone in the past. Therefore, information that may be used to gauge the reputation for honesty and integrity of the others is scant. Working on different aspects or phases of the same project, all agree to share their ideas and findings with the aim of writing joint publications, submitting claims for patents, and drafting joint grant proposals. The prospects for successful collaboration are estimated to be good because each researcher brings to the collaborative effort complementary skills, knowledge, and resources. No formal documents are signed, and complete monitoring is impossible. The collaborative endeavor is based on mutual trust. Our conceptualization of trust in this context is consistent with the definition offered by Rousseau et al. (1998; see also Murphy et al., 2010). Trust is viewed as a cognitive and possibly emotional state of positive expectation (for successful completion of the project by all the participants) in the

face of self-created vulnerability. This expectation is not restricted to a single point in time; rather, it evolves over time.

Given the importance of the project, each group member credited with solving the problem may gain considerable fame, reputation, or money. The longer the collaboration lasts, the higher the value of the joint enterprise. However, the cost of misplaced trust is also potentially high if one of the group members defects and thereby gains the lion's share of the credit. Thus, there is an unavoidable vulnerability that each player must endure. Each group member would like the collaborative effort to continue as its value increases over time. Concurrently, the motivation for betraying the mutual trust increases, too, as the project approaches its termination. Kramer (2001) has referred to these social situations as trust dilemmas, noting that each group member does not have the means to effectively punish the betrayer or reciprocate in any form.

Murphy et al. (2006, 2010) modeled this class of trust dilemmas with a real-time trust game. There are n symmetric players. The strategy space of each player is continuous on the real interval $[0,T]$. Each player can make at most a single decision that terminates the interaction at time $t \in [0, T]$. The game starts at time $t = 0$ and terminates when one of the n players exits the game at some time $t < T$ or when T is reached with no player exiting, whichever occurs first.

Suppose that the game terminates at time $t \in [0, T)$ with a betrayal of trust by player i. Then, the payoff for player i is computed from the exponential payoff function[2] $r_i(t) = \lambda \cdot (2^{(t/\theta)})$ where $\theta \geq 1$ and $\lambda > 0$. The payoff for each of the remaining $n - 1$ players is computed from $r_j(t) = \delta \cdot r_i(t)$ where $0 < \delta < 1$, $j = 1, 2, \ldots, n$, and $j \neq i$. In other words, each of the $n - 1$ players not stopping the clock receives only a fraction δ of the payoff of player i. In continuous time, no tie is possible for $0 < t < T$. If m players ($1 < m \leq n$) stop the clock at exactly $t = 0$, then one of them is randomly chosen with probability $1/m$ to receive the payoff λ, and each of the other $m - 1$ players receives $\delta \cdot \lambda$. If no player stops the clock (and the game terminates at time $t = T$), then the payoff for each of the n players is denoted by g, where $0 \leq g < [\lambda \cdot (2^{(T/\theta)})]$.

The real-time trust game was inspired by the study of Rapoport et al. (2003) on the three-person centipede game and by a subsequent study of Murphy, Rapoport, and Parco (2004) that was designed to investigate the spread of cooperative or noncooperative behavior in a population of participants divided into subsets that iteratively play the three-person centipede game. There are two major differences between the real-time trust game that we review in this chapter and the centipede game (Rosenthal, 1981; Aumann, 1992). First, the real-time trust game is played in continuous time. As a result, the players are symmetric, and the distinction between first mover and second mover disappears. Second, the centipede game provides the players with the opportunity to defect (only once) in sequence and in a rotation that is exogenously determined, while in the real-time trust game there is no exogenous order of play and the players can defect at any time.

The parameters λ and θ control the magnitude and rate of increase in the payoff function, respectively. Together they control greed, which in our model may

enhance motivation but may also at the same time motivate defection. The parameter g controls the incentive to let the clock run without stopping it. Two extreme cases are considered. If g assumes its minimal value, namely, $g = 0$, then the incentive to betray trust (for a given value of δ) is maximized: Each player prefers not joining the collaborative effort rather than letting it terminate with defection. If g assumes its maximal value, namely $g = \lambda \cdot 2^{(T/\theta)}$ and no player stops the clock at time $t = 0$, then no player has an incentive to stop the clock at any time in the game. The parameter δ, which like λ and θ is not conditional on g, controls the incentive to defect but in a different way. Rather than tapping greed, it addresses the fear of having somebody else defect. As δ decreases, the relative difference between the stopper and nonstopper's payoff increases; hence the opportunity cost of misplaced trust can be tuned with this parameter. Conversely, as δ increases and approaches 1, the incentive to defect disappears. The parameters n, T, θ, λ, δ, and g are all commonly known as is the form of the payoff function.

In the equilibrium solution of the game, each player stops the clock at time $t = 0$. This is the case because $r_i(t) > \delta \cdot r_i(t + \varepsilon)$ for any $0 \leq \delta < 1$. Therefore, for any t, it behooves each player to stop the clock before any of the other $n - 1$ players.[3] Continuing the interaction at time t is simultaneously evidence of trust and a weak signal of trustworthiness but not an unequivocal commitment to maintain this behavior for the entire duration of the interaction.

Some psychological factors underlying trust. Greed and fear are two major underlying motives that may impede trust-based cooperation (e.g., Coombs, 1973). Traditionally, they have been studied in the context of games in strategic form including the prisoner's dilemma (Coombs, 1973; Hwang and Burgers, 1997), public good games with provision thresholds (Rapoport and Eshed-Levy, 1989), and more general social dilemmas (Dawes, 1980; Poppe and Utens, 1984). These two motives also underlie behavior in the class of trust dilemmas previously discussed. Greed is the simpler of these motives in that, by definition, it requires only narrow self-interest and no particular beliefs about other players' intentions. A greedy player continues cooperating as long as she believes it is beneficial to do so without regard to the other players' earnings. She defects when she estimates that doing so maximizes her expected payoff. Fear may also motivate a person to defect, if she suspects or anticipates that one of the other players is precariously close to defecting. In continuous-time trust-based dilemma games iterated over time, these preemptive defections may spawn a downward spiral of distrust among otherwise cooperatively minded but fearful players. This is an unfortunate outcome in the sense that the players might have all preferred mutually beneficial outcomes, but in the absence of irrevocable commitments (or some other mechanism) they could not coordinate their joint prosocial preferences to reach collectively beneficial outcomes. Disentangling the impacts of greed and fear is not possible in the present model, but their relative contribution may be changed by manipulating the value of δ. This feature corresponds to earlier theoretical work by Pruitt and Kimmel (1977) and is of use when experimentally exploring the effectiveness of different mechanisms that may foster trust-based cooperation. As Pruitt and Kimmel noted, the

preference for mutual cooperation (e.g., the absence of greed and the presence of prosocial preferences) is not sufficient to yield widespread trust-based cooperation among interacting agents. The preference for mutual cooperation must be accompanied by the belief that one's trustworthy actions will not be taken advantage of or at least that the cost of misplaced trust is not too high. Given the best of intentions, the members of a group of otherwise cooperatively minded players may not take the strategic risk of trusting each other as they fear that such overtures would be met with narrow self-interest. From the greed and fear perspective, we can see that the successful exercise of trust requires both prosocial preferences as well as the belief that other players share similar prosocial preferences.

An example of a real-time trust game. Consider this game with parameter values $T = 45$ (measured in seconds), $\theta = 5$, $\lambda = 5$, $\delta = 0.1$, and $g = 0$; these parameter values provide a clear example of our trust dilemma and also correspond to the parameter values used in one of the conditions of the experiment discussed previously. Thus, if player i stops the clock at some time t, then each of the other $n - 1$ players only receive 10 percent of player i's payoff, whereas player i receives the value of the function at time t. The incentive to defect is particularly strong in this case because $g = 0$. Payoffs are in cents. For any value of $n \geq 2$, if one of the players (for example, player i) stops the clock at time $t \in [0,T]$, then the payoffs rounded to the nearest whole cent (for selected values of t), as presented in table 14.1, are:

If m players stop the clock simultaneously at time $t = 0$, then one of them is chosen with probability $1/m$ to receive the payoff of $\lambda = 5$. Thus, a player may earn between \$0.05, if she stops the clock at exactly $t = 0$ and is chosen to receive λ, and almost \$25.60, if she stops the clock just before 45 seconds.

Figure 14.1 exhibits the payoff function for this example. The figure shows the exponential payoff function that starts at time $t = 0$. Time (on the x-axis) is measured in seconds, and payoff (on the left y-axis) is measured in cents. The player who stops the clock at 40 seconds receives the payoff $5 \cdot (2^{(40/5)}) = \$12.80$. Each of the other two players ($n = 3$ in this example) only receives \$1.28. If the clock were to be stopped by player i at $t = 20$, then player i would have received just \$0.80 and each other player only \$0.08.

We turn next from this example to the general real-time trust game and discuss several of its features. First, conducting noncooperative n-person games in continuous time, although not common in behavioral economics research, is not without precedence. Already in the mid 1970s, several studies reported the results of experiments on a class of two-person zero-sum games known as duels that

Table 14.1 Payoffs for Players in a Real-Time Trust Game

t (in seconds)	0	1	5	10	20	30	35	40	45–ε	45
p_i ("winner")	5	6	10	20	80	320	640	1280	2560–ε	0
p_j *("loser")*	1	1	1	2	8	32	64	128	256–ε	0

Note: Different joint payoffs for the winner and losers are shown for various stopping times.

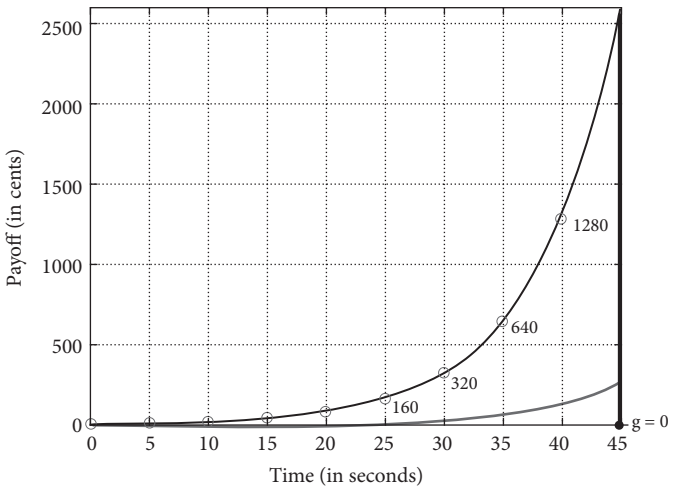

Figure 14.1 Real-time trust game payoff function
Time (x-axis) is measured in seconds, and the winner's payoff (y-axis) is in cents. The top line shows the winner's payoff, whereas the lower line shows the losers' payoffs. The parameter g is indicated showing that in the event that no player stops the clock before T = 45 seconds, all the players earn 0. Some points on the winner's payoff function are highlighted and labeled for clarity.

were modeled and implemented in continuous time. Kahan and Rapoport (1974) reported the results of complete information duels ("noisy duels") with symmetric accuracy functions, and later Kahan and Rapoport (1975) extended the investigation to noisy duels with asymmetric accuracy functions and an asymmetric number of bullets. Rapoport, Kahan, and Stein (1973) further extended these studies to continuous time duels with incomplete information ("silent duels"), and then Rapoport et al. (1976) reported results of probabilistic duels. Almost 30 years later, Kurzban et al. (2001), Goren, Kurzban, and Rapoport (2003), and Goren, Rapoport, and Kurzban (2004) implemented a continuous time protocol of play in which the order of moves and timing of decisions are endogenously determined to study voluntary contributions to the provision of public goods with or without revocable commitments. Interest in continuous time experiments has been revived by the recent study by Oprea, Henwood, and Friedman (2012) of the standard hawk-dove game and their new software package called ConG that allows the players to make asynchronous decisions in continuous time, receive instantaneous feedback, and change their decisions as often as they wish.

Second, the basic game may be generalized in several different ways. The exponential payoff function (also used by Rapoport et al., 2003, and Murphy et al., 2004) may be replaced by any other monotonically increasing function—a linear function would be a natural alternative (see, as an example, Murphy, 2010). The parameter δ may be set to some positive value, so that all the players earn a positive

amount if they exercise mutual trust. The parameter δ may be individualized in order to introduce asymmetry between the players. The parameter T, rather than being fixed and known, may be replaced with a random distribution function over T that is commonly known to the players.

Third, alternative methods may be used to elicit decisions from the players based on how the basic game is implemented. Under the decision method, only a single stopping time is recorded when player i decides to stop the clock. The remaining $n - 1$ players are not presented with the opportunity to register their intended stopping times as the game has already been stopped while they were waiting. Therefore, under this method it is not possible to elicit information about the propensity to cooperate by the (necessarily more trusting) $n - 1$ players. Conversely, under the strategy method, all the n players register their stopping times at time $t = 0$ independently of the other players, and the game always progresses until time T. Only at time T is each of the n players informed about the decisions of the group members, and payoffs are determined just as before as a function of the minimum stopping time. The strategy method facilitates credible signaling (if in the iterated basic game the stopping times are rendered common knowledge) and yields researchers a richer body of data for analysis.

Fourth, collaborative situations of the type modeled by the real-time trust game have clear parallels outside the laboratory. Vanderkam and Flint (2002) describe a situation in which mutual trust that had been maintained for 40 years broke down among several distinguished scholars who had undertaken the task of studying, interpreting, and jointly publishing fragments of the Dead Sea scrolls that had been discovered in the Judean desert between 1947 and 1956 and distributed for analysis and interpretation. The breakdown of mutual trust resulted in long delays in publication, considerable professional friction, and ultimately protracted lawsuits. All of these researchers would have been better off trusting one another and acting in a trustworthy manner. However, each had an incentive to act in an untrustworthy way, and some combination of fear and greed unraveled an otherwise valuable professional opportunity.

Two illustrative studies

Effects of Group Size and the Opportunity Cost

We know from the study of the n-person prisoner's dilemma game ($n \geq 2$) that the actual values of the payoffs and the number of players affect the level of cooperation. Players tend to defect more often as the temptation to defect increases (e.g., Rapoport and Chammah, 1965; Bonacich et al., 1976), and the level of cooperation drastically decreases as the number of players increases. Murphy et al. (2006) designed a preliminary study of the real-time trust game to test these two

directional hypotheses. For this purpose, they had 126 subjects participating in six independent sessions, each including 21 players. Four of the six parameters of the game were fixed across sessions, namely, $T = 45$ seconds, $\theta = 5$, $\lambda = 5$, and $g = 0$. The other parameter values were varied across three conditions:

Condition $n = 3/\delta = 0.5$	Relatively small opportunity cost	Sessions 1 and 2
Condition $n = 3/\delta = 0.1$	Baseline	Sessions 3 and 4
Condition $n = 7/\delta = 0.1$	Relatively large group size	Sessions 5 and 6

Each condition was replicated twice for a total of six sessions. Sessions 1 and 2 included seven groups of three players each, as did sessions 3 and 4. Sessions 5 and 6 included three groups of seven players each. Within each session, the basic game was iterated for 90 rounds of play. Group membership was randomly assigned on each round to minimize sequential dependencies and prevent reputation building. Communication among group members was not possible. The decision method was used in each session. Murphy et al. (2006) reported four major findings.

Finding 1. Mean stopping times in condition $n = 3/\delta = 0.5$ decreased steadily from about 35 seconds on round 1 to about 30 seconds on round 90. Small as it may appear, this five-second difference translates to about 50 percent drop in the payoff for the player stopping the clock, from \$6.40 to \$3.20. We interpret this finding as evidence for the gradual degradation of trust-based cooperation in the population in which neither reputation building nor punishment for defection is allowed. Mean stopping time in condition $n = 3/\delta = 0.1$ followed a similar pattern, although the change over time was more dramatic. Median stopping times started at about 31 seconds on round 1 and dropped to 15 seconds by round 90. This translates to a considerable reduction in mean payoff across the 90 iterations of the stage game, from about \$3.20 to a meager \$0.40. The breakdown in mutual trust was further accelerated in condition $n = 7/\delta = 0.1$; the median stopping times on round 1 was about 17 seconds, and in both sessions 5 and 6 it converged to zero rapidly. To provide perspective, this dismal result entails a population of players competing vigorously over an average expected payoff of about 3 cents per round; had they all trusted and cooperated with one another by stopping the clock after, for example, 40 seconds they all would have earned on average more than 100 times this amount.

Finding 2. Because of the random assignment of subjects to groups, a strong social norm was established in the entire population dictating the time interval to stop the clock. Although median stopping time decreased in all six sessions, the size of this interval remained more or less constant. It is worth noting that the strong conformity to this norm (independent groups each stopping the clock within about four to five seconds of one another) was developed in the absence of any agreements or communication between players.

Finding 3. The results support the hypothesis that decreasing the value of δ increases the temptation to defect. The null hypothesis that the median stopping

time in condition $n = 3/\delta = 0.5$ is equal to that in condition $n = 3/\delta = 0.1$ was soundly rejected (Mann-Whitney test, $z = -15.35$, $p < 0.001$). Toward the end of the session, subjects in condition $n = 3/\delta = 0.5$ were earning about eight times more than subjects in condition $n = 3/\delta = 0.1$.

Finding 4. The results also support the hypothesis that as group size increases the temptation to defect earlier increases too. The null hypothesis that the median stopping time in condition $n = 3/\delta = 0.1$ is equal to the median stopping time in condition $n = 7/\delta = 0.1$ was soundly rejected (Mann-Whitney test, $z = 14.13$, $p < 0.001$).

Effects of Credible Signaling

A general finding from social dilemma research is that rates of cooperation decrease over repeated interactions. This gloomy finding (Ostrom, 2003) has resulted in a variety of research streams testing mechanisms that could potentially stave off the unraveling of trust-based cooperation. Communication (Deutsch, 1960), reputation (Berg, Dickhaut, and McCabe, 1995), and punishment (Fehr and Gächter, 2002) are the better-known mechanisms that have been found to be useful in sustaining trust. We explored the effects of a different mechanism, namely credible signaling[4], as a means of fostering cooperation. This mechanism has advantages as it is noncentralized, anonymity preserving, and efficient. It is noncentralized in the sense that no central authority or monitoring agent is required for its implementation, as is the case for reputation. Credible signaling can also be implemented in anonymous interactions, whereas face-to-face communication cannot; reputation, too, is inconsistent with anonymity. By efficiency, we mean that value is not destroyed by the mechanism, as is the case with punishment. Although punishment has a clear effect on cooperation rates of interdependent decision makers over time (Fehr and Gächter, 2002), the resulting aggregate earnings are not concordantly better because the cooperation dividend is spent on enacting costly punishments, thus destroying value.

Murphy et al. (2010) designed and conducted experiments to test the effects of credible signaling on the dynamics of trust evolution. Central to their design was the use of the strategy method that allowed all players to simultaneously register their intended stopping times. Although the payoffs in the game are by definition exclusively a function of one player's stopping time (the player who stopped the clock first), the information contained in the choices of the other $n - 1$ players serve as credible signals of genuine cooperative intent, especially if these signals persist over multiple iterations. These signals can serve to mitigate the fear among players of their cooperative moves being taken advantage of.[5] The results show a substantial effect of signaling when compared to the baseline condition from Murphy et al. (2006) for which signaling was impossible. There are two major results of note. First, the enduring choices of trusting and trustworthiness (as manifested by letting the clock run just short of T) from several "hard-core cooperators" percolated throughout the population of players and led to significantly greater earnings for all the players. There is evidence that a few "good apples" can serve as a rallying point

for other cooperatively minded but otherwise fearful or greedy players, thus facilitating a norm of trust-based cooperation emerging in the population of anonymous players. Second, the downward trend of unraveling trust, which is commonly found in social dilemma research, was not observed in this experiment. Rather, the population of players reached a steady state of part-way trust-based cooperation, with winners stopping the clock at about 25 seconds on average. By the end of the experimental session this is about ten seconds longer than their counterparts in the baseline condition who were not granted the option of credible signaling. This difference in stopping time corresponds to a fourfold increase in mean earnings, an effect that is both statistically significant and nontrivial.

Conclusion

This chapter describes and illustrates a simple game designed to experimentally test the evolution and breakdown of cooperative behavior that is based on mutual trust. The real-time trust game is particularly suitable to study the dynamics related to the emergence, maintenance, and potential unraveling of trust-based cooperation. Preliminary results reported in part 3 show that the game elicits patterns of behavior that are fully consistent with patterns previously reported in the literature: The potential opportunity cost of betrayed trust matters, group size matters, and the overall level of trust in the population decreases as the stage game it iterated over time. Moreover, the real-time trust game has been used to study the effects of credible signaling in the absence of reputation or punishment, and this inexpensive mechanism has been identified as an effective means of preserving a moderate degree of trust-based cooperation.

We have proposed this game in order to study elements of trust that are presently not captured by standard, single-shot, two-person extensive-form games. Generalizations of the basic game may be proposed to address other variables that affect the development and breakdown of trust-based cooperative behavior. We already have mentioned a model according to which the duration of the interaction, rather than being finite and known, is a random variable with a known distribution function. Alternatively, the interaction may be allowed to continue indefinitely with a fixed probability of an exogenous breakdown at any time during the game. In another generalization of the basic game, the degree of the temptation to defect may be individualized. The effects of "hard-core cooperators" who were identified in the study of Murphy et al. (2010) may further be studied by introducing "bots" that are programmed to stop the clock just short of T or not stop it at all. Noticing that many alliances, both economic and political, may sustain the defection of a single member but collapse if a subset of players defects, the basic model may further be generalized by requiring that the game terminates only after k members ($1 < k \leq n$) defect to collect their individual rewards.

Acknowledgment

This work was supported by NSF grant SES-0637151 to Columbia University and the University of Arizona. We wish to thank Vincent Mak for a critical reading of the manuscript and very helpful comments.

Notes

1. The real-time trust game that we propose below differs from public good games by endowing each of the *n* players with veto power. In most but not all public good games, the public good may be produced even if some of the group members defect. In our model, all the *n* players have to sustain cooperation over time in order for the collective gain to be fully realized. Our model is more akin to the volunteer's dilemma (see, e.g., Diekmann, 1985, 1993; Weesie, 1993, 1994; and, in particular, the class of dynamic volunteer's dilemma games proposed by Otsubo and Rapoport, 2008).
2. We use here a base 2 exponential payoff function. However, any monotonically increasing function would also be appropriate. A linear function would be a natural alternative (see the trust allocation game in Murphy, 2010, as an example). The payoff structure used here is isomorphic to that used in previous experimental centipede games (Rapoport et al., 2003).
3. Stated informally, in the subgame at time $t = T - \varepsilon$, it is a strictly dominant strategy to stop, and therefore all the *n* players should stop at time $t = T - \varepsilon$. Repeated backward induction over ε intervals, equivalent to iterated elimination of dominated strategies, leads to the equilibrium of stopping at time $t = 0$.
4. A credible signal is different from cheap talk in that the former is costly for a signaler to use and thus demonstrates a real and reliable motivation on their part for cooperation and the subsequent realization of joint gain. See Farrell (1993) for more details.
5. Of course, signaling the intention to cooperate in future rounds by stopping the clock on round *h* just short of *T* is still ambiguous. A player may anonymously record a late stopping time just to dupe other players into stopping late in subsequent rounds, even if that means sacrificing a payoff in round *h*. Randomly dividing the population into small groups on each round—as in the present study—considerably reduces this ambiguity. To ensure that such signals are fully credible, the identities of the players who record relatively late stopping times may have to be revealed.

References

Aumann, R. J. (1992). Irrationality in game theory. In P. Dasgupta, D. Gale, O. Hart, and E. Maskin (eds.), *Economic Analysis of Markets and Games: Essays in Honor of Frank Hahn*, 214–227 Cambridge, MA: MIT Press.

Bacharach, M., Guerra, G., and Zizzo, D. J. (2007). Is trust self-fulfilling? an experimental study. *Theory and Decision,* 63(4), 349–388.

Berg, J., Dickhaut, J., and McCabe K. A. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior*, 10, 122–142.

Bonacich, P., Shure, G., Kahan, J., and Meeker, R. (1976). Cooperation and group size in the n-person prisoners' dilemma. *Journal of Conflict Resolution*, 20, 687–706.

Burnham, T., McCabe, K. A., and Smith, V. L. (2000). Friend-or-foe intentionality priming in an extensive form trust game. *Journal of Economic Behavior and Organization*, 43, 57–73.

Butler, J. V., Giuliano, P., and Guiso, L. (2010). Cheating in the trust game. European University Institute, EIEF, and CEPR, unpublished manuscript.

Camerer, C. F. (2003). *Behavioral Game Theory: Experiments in Strategic Interaction*. New York: Russell Sage Foundation.

Cardenas, J. C. and Carpenter, J. P. (2008). Behavioral development economics: lessons from field labs in the developing world. *Journal of Development Studies*, 44, 337–364.

Coombs, C. H. (1973). A reparameterization of the Prisoner's Dilemma game. *Behavioral Science*, 18, 424–428.

Cox, J. C. (2002). Trust, reciprocity, and other-regarding preferences: groups vs. individuals and males vs. females. In R. Zwick and A. Rapoport (eds.), *Experimental Business Research,* 331–349. New York: Kluwer.

Dasgupta, P. (1988). Trust as a commodity. In D. Gambetta (ed.), *Trust: Making and Breaking Cooperative Relations*. New York: Basil Blackwell.

Dawes, R. M. (1980). Social dilemmas. *Annual Review of Psychology*, *31*, 169–193.

Deutsch, M. (1960). The effect of motivational orientation upon trust and suspicion. *Human Relations*, 13, 123–139.

Diekmann, A. (1985). Volunteer's dilemma. *Journal of Conflict Resolution*, 29, 611–618.

_____. (1993). Cooperation in an asymmetric volunteer's dilemma: theory and experimental evidence. *International Journal of Game Theory*, *22*, 75–85.

Engle-Warnick, J. and Slonim, R. L. (2001). The fragility and robustness of trust. Case Western Reserve University, Department of Economics, unpublished manuscript.

Farrell, J. (1993). Meaning and credibility in cheap-talk games. *Games and Economic Behavior*, 5, 514–531.

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

Fukuyama, F. (1995). *Trust: The Social Virtues and the Creation of Prosperity*. New York: Free Press.

Glaeser, E. L., Laibson, D. L., Scheinkman, J. A., and Soutter, C. L. (2000). Measuring trust. *Quarterly Journal of Economics*, 125, 811–846.

Goren, H., Kurzban, R., and Rapoport, A. (2003). Social loafing vs. social enhancement: public goods provisioning in real-time with irrevocable commitments. *Organizational Behavior and Human Decision Processes*, 90, 277–290.

Goren, H., Rapoport, A., and Kurzban, R. (2004). Revocable commitments to public goods provision under the real-time protocol of play. *Journal of Behavioral Decision Making*, 17, 17–37.

Güth, W., and Kliemt, H. (1994). Competition or co-operation—on the evolutionary economics of trust, exploitation and moral attitudes. *Metroeconomica*, 45, 155–187.

Güth, W., Ockenfels, A., and Wendel, M. (1993). Efficiency by trust or fairness? multiperiod ultimatum bargaining experiments with an increasing cake. *International Journal of Game Theory*, 22, 51–73.

Güth, W., Ockenfels, A., and Wendel, M. (1997). Cooperation based on trust: an experimental investigation. *Journal of Economic Psychology*, 18, 15–43.

Hardin, R. (2004). *Trust and Trustworthiness*. New York: Russell Sage Foundation.

Ho, T. H. and Weigelt, K. (2005). Trust building among strangers. *Management Science,* 51(4), 519–530.

Hwang, P. and Burgers, W. P. (1997). Properties of trust: an analytical view. *Organizational Behavior and Human Decision Processes*, 69, 67–73.

Kahan, J. P. and Rapoport, A. (1974). Decisions of timing in bipolarized conflict situations with complete information. *Acta Psychologica*, 38, 183–203.

_____. (1975). Decisions of timing in conflict situations of unequal power between opponents. *Journal of Conflict Resolution*, 19, 250–270.

Kramer, R. M. (2001). Trust rules for trust dilemmas: how decision makers think and act in the shadow of doubt. In R. Falcone, M. Singh, and Y.-H. Tan (eds.), *Trust in Cyber-societies,* 9–26. Berlin: Springer-Verlag.

Kurzban, R., McCabe, K. A., Smith, V. L., and Wilson, B. (2001). Incremental commitment and reciprocity in a real-time public goods game. *Personality and Social Psychology Bulletin*, 27, 12, 1662–1673.

Kurzban, R., Rigdon, M. and Wilson, B. (2008). Incremental approaches to establishing trust. *Experimental Economics*, 11, 4, 370–389.

McCabe, K. A., Rassenti, S. J., and Smith, V. L. (1996). Game theory and reciprocity in some extensive form experimental games. *Proceedings of the National Academy of Sciences*, 93, 13421–13428.

McCabe, K. A., Rassenti, S. J., and Smith, V. L. (1998). Reciprocity, trust, and payoff privacy in extensive form bargaining. *Games and Economic Behavior*, 24, 10–24.

McCabe, K. A., Rigdon, M., and Smith, V. L. (2002). Cooperation in single play, two-person extensive form games between anonymously matched players. In R. Zwick and A. Rapoport (eds.), *Experimental Business Research,* 51–67. New York: Kluwer.

McCabe, K. A., Smith V. L., and LePore, M. (2000). Intentionality detection and "mindreading:" why does game form matter? *Proceedings of the National Academy of Sciences*, 97, 4404–4409.

Murphy, R. O. (2010). The trust allocation game. The Chair of Decision Theory and Behavioral Game Theory, ETH Zürich, unpublished manuscript.

Murphy, R. O., Rapoport, A., and Parco, J. E. (2004). Population learning of cooperative behavior in a three-person centipede game. *Rationality and Society*, 16, 91–120.

_____. (2006). The breakdown of cooperation in iterative real-time trust dilemmas. *Experimental Economics*, 9, 147–166.

_____. (2010). Credible signaling in real-time trust dilemmas. The Chair of Decision Theory and Behavioral Game Theory, ETH Zürich, unpublished manuscript.

Ockenfels, A. and Resnick, P. (2012). Negotiating reputations. In G. Bolton and R. Croson (Eds.), *Oxford Handbook of Economic Conflict Resolution*. Oxford, UK: Oxford University Press.

Oprea, R., Henwood, K., and Friedman, D. (2012). Separating the hawks from the doves: evidence from continuous time laboratory games. *Journal of Economic Theory*.

Ortmann, A., Fitzgerald, J., and Boeing, C. (2000). Trust, reciprocity, and social history: a re-examination. *Experimental Economics*, 3, 81–100.

Ostrom, E. (2003). Toward a behavioral theory linking trust, reciprocity, and reputation. In E. Ostrom and J. Walker (eds.), *Trust and Reciprocity,* 19–79. New York: Russell Sage Foundation.

Otsubo, H. and Rapoport, A. (2008). Dynamic volunteer's dilemmas over a finite horizon: an experimental study. *Journal of Conflict Resolution*, 52, 961–984.

Poppe, M. and Utens, L. (1984). Effects of greed and fear of being gypped in a social dilemma situation with changing pool size. *Journal of Economic Psychology*, 7, 61–73.

Pruitt, D. G. and Kimmel, M. J. (1997). Twenty years of experimental gaming: critique, synthesis, and suggestions for the future. *Annual Review of Psychology*, 28, 363–392.

Raiffa, H. (2002). *Negotiation Analysis*. Cambridge, MA: Belknap Press of Harvard University Press.

Rapoport, Anatol, and Chammah, A. M. (1965). *Prisoner's Dilemma A Study in Conflict and Cooperation*. Ann Arbor: University of Michigan Press.

Rapoport, A. and Eshed-Levy, D. (1989). Provision of step-level public goods: effects of greed and fear of being gypped. *Organizational Behavior and Human Decision Processes*, 44, 325–344.

Rapoport, A., Kahan, J. P., and Stein, W. E. (1973). Decisions of timing in conflict situations of incomplete information. *Behavioral Science*, 18, 272–287.

Rapoport, A., Kahan, J. P., and Stein, W. E. (1976). Decision of timing in experimental probabilistic duels. *Journal of Mathematical Psychology*, 13, 163–191.

Rapoport, A., Stein, W. E., Parco, J. E., and Nicholas, T. E. (2003). Equilibrium play and adaptive learning in a three-person centipede game. *Games and Economic Behavior*, *43*, 239–265.

Resnick, P. and Zeckhauser, R. (2002). Trust among strangers in Internet transactions: empirical analysis of eBay's reputation system. In M R. Baye (ed.), *The Economics of the Internet and E-Commerce*, 127–157. New York: JAI Press.

Rosenthal, R. W. (1981). Games of perfect information, predatory pricing and the chain-store paradox. *Journal of Economic Theory*, 25, 92–100.

Rousseau, D., Sitkin, S., Burt, R., and Camerer, C. (1998). Not so different after all: a cross-discipline view of trust, *Academy of Management Review*, 23, 3, 393–404.

Vanderkam, J. C., and Flint, P. (2002). *The Meaning of the Dead Sea Scrolls*. New York: Harper.

Weesie, J. (1993). Asymmetry and timing in the volunteer's dilemma. *Journal of Conflict Resolution*, 37, 569–590.

Weesie, J. (1994). Incomplete information and timing in the volunteer's dilemma: a comparison of four models. *Journal of Conflict Resolution*, 38, 557–587.