Exp Econ (2006) 9: 147–166
DOI 10.1007/s10683-006-7049-4

The breakdown of cooperation in iterative real-time trust dilemmas

Ryan O. Murphy · Amnon Rapoport · James E. Parco*

Received: 20 September 2005 / Revised: 14 November 2005 / Accepted: 14 November 2005

Abstract We study a class of trust-based cooperation dilemmas that evolve in continuous time. Characteristic of these dilemmas is that as long as all n players continue to cooperate, their payoffs increase monotonically over time. Simultaneously, the temptation to defect increases too, as the first player to defect terminates the interaction and receives the present value of the payoff function whereas each of the other $n - 1$ players only receives a proportion δ ($0 < \delta < 1$) of the defecting player's payoff. We introduce a novel experimental institution that we call the Real-Time Trust Game (RTTG) to examine this class of interactions. We then report the results from an iterated RTTG in which the values of n and δ are varied in a between-subjects design. In all conditions, cooperation breaks down in the population over iterations of the game. The rate of breakdown sharply increases as n increases and more slowly decreases as δ increases.

Keywords Centipede game · Cooperation · Population dynamics · Trust · Trust dilemmas

JEL Classification C72 · C92 · D81

*The thoughts and opinions expressed in this manuscript are solely those of the authors and do not necessarily reflect the views of the U.S. Government, the Department of Defense, or the United States Air Force.

R. O. Murphy (✉)
Columbia University, Center for the Decision Sciences, 420 West 118th Street, Room 805A Mail Code 3355, New York, NY 10027
e-mail: rom2102@columbia.edu

A. Rapoport
University of Arizona, Dept. of Management and Policy, 405 McClelland Hall, Tucson, AZ 85721 and Hong Kong University of Science and Technology, Dept. of Economics, Kowloon, Hong Kong
e-mail: amnon@u.arizona.edu

J. E. Parco
United States Air Force Academy, Dept. of Management, Colorado Springs, CO 80840
e-mail: james.parco@usafa.edu

Introduction

Trust dilemmas in real time

We focus on n-person interactive situations that evolve over time in which players are symmetric, cooperation is based on mutual trust, no monitoring is possible, the joint fruits of cooperation increase over time, and any player can unilaterally terminate the interaction at any time. To illustrate these situations, consider a hypothetical vignette similar to Kramer (2001) of three researchers who work on a problem of shared interest and decide to enter into scientific collaboration. None of the researchers know a great deal about the others, as all have worked mostly alone in the past. Thus, the information that can be used to assess the trustworthiness of the others is scant. Each of the researchers is working on a problem of considerable scientific interest and practical importance, and all three agree to share their ideas and findings with the aim of writing joint publications, placing claims for a patent, and so on. The prospects for a successful and fruitful collaboration are very promising because each researcher brings to the collaboration complementary resources, skills, or knowledge. As is common in such collaborative efforts, no formal documents are signed and no monitoring is possible. The cooperative endeavor is sustained by trust.

Given the importance of the problem, any member of the team credited with solving the problem under investigation receives considerable claim, reputation, or monetary gain. The longer the collaboration lasts, the higher the value of the joint enterprise (to use a term from the study of integrative bargaining, the "pie increases in size"). The cost of misplaced trust is potentially high, if one of the researchers "defects" from the collaboration thereby garnering the lion's share of the credit. Each member would like the collaborative effort to continue as its value increases in time. Concurrently the motivation for betraying trust increases too.

Such situations illustrate a broad class of interactive decision-making problems that Kramer calls *trust dilemmas*. In trust dilemmas agents interact with one another over time, and each hopes to reap some benefit from engaging in cooperative behavior based on trust. The longer players continue their collaborative effort, the higher the joint benefit. Pursuit of this opportunity exposes each agent to the prospect that his or her trust might be exploited. Once betrayed, he or she cannot punish the betrayer or reciprocate in any other form. Kramer claims "Because of our dependence on, and interdependence with, other social decision-makers, trust dilemmas are inescapable feature of social and organizational life (2001, p. 10)."

Experimental trust games

Experimental studies of trust and trustworthiness, that originated with the pioneering works of Güth and Kliemt (1994) and Berg et al. (1995) have typically created experimental settings where there are no institutional mechanisms conducive to trusting and trustworthy behavior. Most of the experimental institutions have provided minimal scope for personal relations and social networks by randomly sampling and matching subjects, conducting the experiments under complete anonymity, and eliminating other design features (e.g., repeated games with fixed group assignment) that could, in part, sustain cooperative behavior without trust (e.g., as in the finitely iterated Prisoner's Dilemma game). These experimental settings have yielded consistent and replicable evidence that many, although not all, subjects do not follow self-interest dominant pure strategies, nor do they expect such behavior from their counterparts. Principles of dominance and backward induction, which play a critical role in noncooperative game theory, fail to account for what seems to be the cooperative behavior of these subjects. For a representative sample of experiments on trust and reciprocity, (see Bacharach et al., 2001; Burnham et al., 2000; Camerer and Weigelt 1988; Cox, 2002; Glaeser et al., 2000; Güth et al., 1993, 1997; Ho and Weigelt, 2001; McCabe et al., 2002; McCabe et al., 1996, 1998; McCabe et al., 2000; Ortmann et al., 2000; Engle-Warnick and Slonim 2001 and Rapoport et al., 2003). Camerer (2003) provides a comprehensive literature review.

The centipede game

Of particular relevance to our study are extensive form games of the centipede type in which trust has been invoked to account for deviations from equilibrium play. A two-person centipede game was introduced by Rosenthal (1981) and later studied theoretically by Aumann (1992, 1995, 1998), Ben-Porath (1997), Feinberg (2001), Ponti (2000), Reny (1993), Stalnaker (1996, 1998), and many others. Figure 1 displays a variant of the centipede game discussed by Aumann (1992). There are two players in this game called Alice and Ben, and a sum of \$10.50 lying on the table in front of them. Moving first, Alice has the option of taking \$10.00, leaving \$0.50 to Ben. If she chooses this option ("exit"), the game is over. If not ("continue"), the amount on the table is increased tenfold, and it is Ben's turn to play next. He now has the option of taking \$100.00, leaving \$5.00 to Alice. If he does so ("exit"), the game is over. If not, ("continue"), the joint amount is increased again tenfold to \$1,050. Continuing in this way, with player roles being interchanged and payoff increasing tenfold on each move (stage), the game terminates after three full rounds of play (i.e., six moves). In the final stage, Ben has the option of taking \$1,000,000, leaving Alice \$50,000. If he continues, the game terminates with each player receiving nothing. As shown in Fig. 1, the game has six decision nodes and seven terminal nodes. Associated with each terminal node are two numbers: the top number is Alice's payoff and the bottom number is Ben's payoff.

This game has two central features that are shared by other centipede games. First, the payoffs are structured in such a way that both players are better off if play continues for at least two stages, except for the last stage. Second, if one player continues and the other exits on the immediately following stage, then the player who continued is worse off and the one who exited is better off. Therefore, at each stage there is an incentive to defect (exit) rather than cooperate (continue) and thereby risk a smaller payoff (Rapoport, 2003). Each of the games in Fig. 1 is a finite extensive form with perfect information. The unique solution of such games, obtained by backward induction, is to exit on each stage. This solution is implied by the assumption of common knowledge of material rationality (Aumann, 1998). Many reasonable people are unwilling to accept this solution, or at least believe that it represents an approach of little practical value (Aumann, 1992) in this context.

The critical question is what to make out of Alice's decision if, deviating from equilibrium play, she decides to continue on her first move. Ben may interpret this move in one of several possible ways. He may ascribe her move to an error, possibly based on her misunderstanding of the payoff function or the nature of the game. Alternatively, he may believe that Alice fully understands the game and is, in fact, trying to trick him to continue on *his* first move so that she can exit immediately on *her* second move and increase her payoff one-hundred fold. Yet another interpretation that Ben may consider is that Alice's decision to continue on her first move is a signal of her intention to cooperate by continuing beyond her second move so that, regardless of who exits, *both* Alice and Ben will gain higher payoffs than those associated with an exit on Alice's first move. This interpretation invokes the notion of trust. When deciding to continue on her first decision node, Alice has to trust that Ben will not exit on his first decision node and cause her to end up with \$5.00 rather than the \$10.00 she could have had. As James (2002) writes, when we say that "A trusts B" we typically mean that A expects B not to exploit a vulnerability that A created by taking an action. Berg

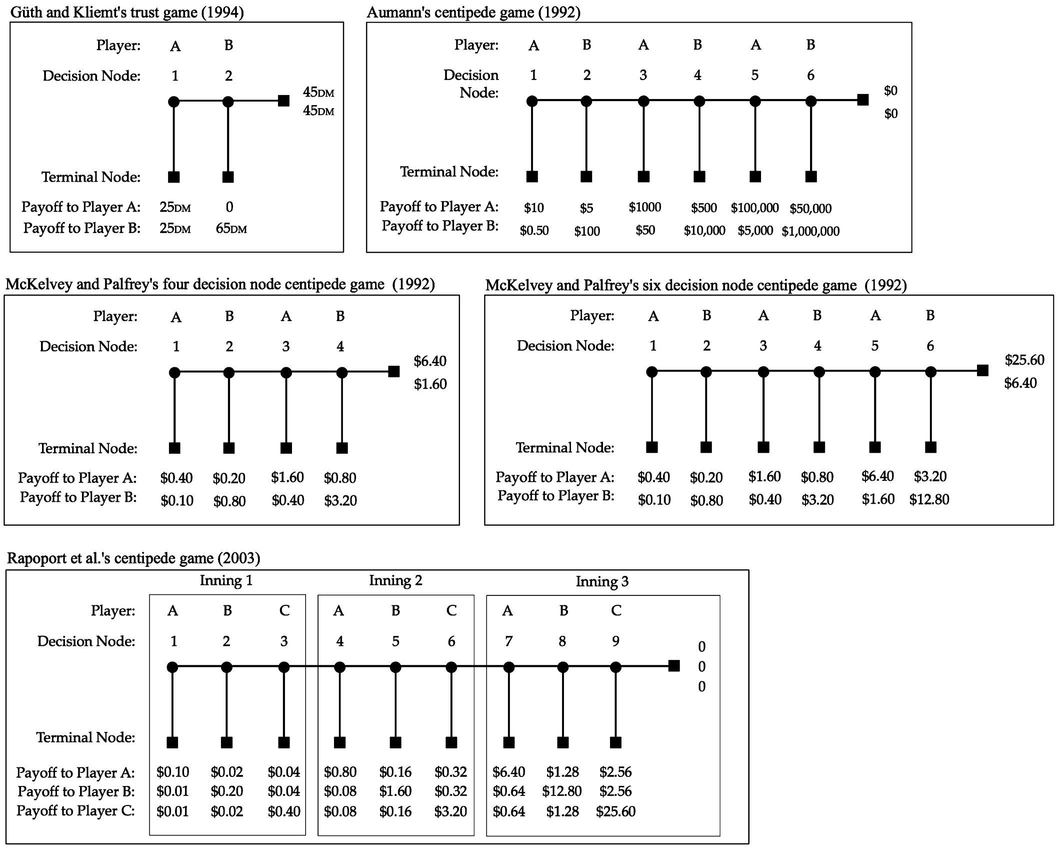

Güth and Kliemt's trust game (1994)
Player: A B
Decision Node: 1 2
45DM
45DM
Terminal Node:
Payoff to Player A: 25DM 0
Payoff to Player B: 25DM 65DM
Aumann's centipede game (1992)
Player: A B A B A B
Decision Node: 1 2 3 4 5 6
$0
$0
Terminal Node:
Payoff to Player A: $10 $5 $1000 $500 $100,000 $50,000
Payoff to Player B: $0.50 $100 $50 $10,000 $5,000 $1,000,000
McKelvey and Palfrey's four decision node centipede game (1992)
Player: A B A B
Decision Node: 1 2 3 4
$6.40
$1.60
Terminal Node:
Payoff to Player A: $0.40 $0.20 $1.60 $0.80
Payoff to Player B: $0.10 $0.80 $0.40 $3.20
McKelvey and Palfrey's six decision node centipede game (1992)
Player: A B A B A B
Decision Node: 1 2 3 4 5 6
$25.60
$6.40
Terminal Node:
Payoff to Player A: $0.40 $0.20 $1.60 $0.80 $6.40 $3.20
Payoff to Player B: $0.10 $0.80 $0.40 $3.20 $1.60 $12.80
Rapoport et al.'s centipede game (2003)
Inning 1
Inning 2
Inning 3
Player: A B C A B C A B C
Decision Node: 1 2 3 4 5 6 7 8 9
0
0
0
Terminal Node:
Payoff to Player A: $0.10 $0.02 $0.04 $0.80 $0.16 $0.32 $6.40 $1.28 $2.56
Payoff to Player B: $0.01 $0.20 $0.04 $0.08 $1.60 $0.32 $0.64 $12.80 $2.56
Payoff to Player C: $0.01 $0.02 $0.40 $0.08 $0.16 $3.20 $0.64 $1.28 $25.60

Fig. 1 Examples of extensive form trust games

et al. (1995) were the first to notice that their two-stage investment game, designed to study trust and reciprocity, and the two-person centipede game invoke the same notion of trust. They write "The centipede game may go on for many stages, but any two consecutive stages involve the same basic structure," and later "The investment game proposed in this paper provides a 'boundary' design for the centipede game" (p. 123). The basic difference between the investment game and the centipede game is that the latter has only two alternatives at each decision node, whereas the former has a larger strategy space allowing different degrees of trust.

Experimental studies of the centipede game have been conducted by McKelvey and Palfrey (1992), Fey et al. (1996), Nagel and Tang (1998), Ho and Weigelt (2001), Rapoport et al. (2003), all designed in part to assess the descriptive power of the equilibrium solution. Two experiments by McKelvey and Palfrey and Experiment 2 by Rapoport et al. are most relevant to the present study (see Fig. 1 for representations). Despite the differences between them in the number of players ($n = 2$ vs. $n = 3$), payoffs associated with the decision to continue on the final decision node (substantial payoffs that maximize joint outcome vs. zero), number of iterations of the stage game (10 vs. 60), and the matching protocol (each player is matched with each other player exactly once vs. random assignment of players to groups and player roles within groups on each round), these experiments yielded very similar results. First and most importantly, neither study's results supported equilibrium play. Second, the majority of the interactions terminated in the middle part of the game tree, indicating some level of trust was generally exercised. Third, in all the experiments there was either no evidence or only very weak evidence for learning across iterations of the stage game.

Limiting features of extensive form trust games

Several limiting features of previous experiments on trust are noted. First, with the exception of the study by Rapoport et al. (2003), previous trust studies 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. Organizations often employ task specific teams that require trust and cooperation in order to be effective. Members of economic alliances have to trust one another to perform their share of the joint project when perfect monitoring is not possible. In these examples trust and cooperation may be manifested in groups with more that two players.

Second, all previous experiments that were mentioned above are restricted to a discrete strategy space. Experimental paradigms with discrete strategy space certainly have their advantages. It is very convenient to portray interactions as extensive form games and construct theories (e.g., subgame perfection and other equilibrium refinements) for such games. It is also easy to implement extensive form games in laboratory experiments. But as mentioned earlier, many, if not most, social interactions, particularly those taking place in populations rather than dyads, evolve steadily in real time.

A third feature of these previous studies is that they restrict the number of actions (decision nodes) available to the players. In principle, a game tree is not limited in size, and experimenters could add decision nodes indefinitely, extending the game tree at will. But the documented difficulties that people encounter with backward induction (e.g., in the Rubinstein sequential bargaining game [1982]) impose practical constraints on the total number of decision nodes that can be plausibly implemented. Further, representations of games with more than about 10 decision nodes can be cumbersome and difficult to represent on a computer screen during an experiment. The game that we propose below has been devised to circumvent this technical limitation.

A fourth and arguably most critical limitation of the experimental trust institutions used to date is the built-in asymmetry between the players. Extensive form games necessarily unfold sequentially with players making their choices in turn. As a consequence, players have different roles that may then influence their choices. Asymmetries can deny particular players the opportunity to register *any* move. Consider all the games exhibited in extensive form in Fig. 1. In equilibrium, each of these games terminates with the first mover choosing to exit, thereby providing no opportunity to the other players to participate. The trust games mentioned earlier, where one player may offer to the second player any fraction of her endowment, are also plagued by this same problem where in equilibrium one of the two members of the dyad may not be called upon to act at all. Ideally, one would like to have a game in which the players are unencumbered by exogenously defined roles.

The real time trust game

We model a class of trust dilemmas with the following real-time trust game (RTTG) that overcomes these limitations. 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 "stops the clock" at time $t \in [0,T)$. The game starts at time $t = 0$ and terminates either when one of the n players (called the "winner") stops the clock (or "exits") at some time $t < T$ or when T is reached with no player stopping the clock, whichever occurs first.

Suppose that the game terminates at time $t \in [0,T)$ with player i stopping the clock. Then, the payoff for the (single) winner i is computed from the exponential payoff function[1] $r_i(t) = \lambda \times (2^{(t/\theta)})$ where $\theta \geq 1$ and $\lambda > 0$. The payoff for each of the remaining $n - 1$ "losers" is computed from $r_j(t) = \delta \times r_i(t)$ where $0 < \delta < 1$, $j = 1, 2, \ldots, n$, and $j \neq i$. In words, each of the $n-1$ players not stopping the clock receives the fraction δ of the winner's payoff.

As time is continuous, no tie is possible at times $0 < \mathrm{t} < T$. If m players $(1 < m \leq n)$ stop the clock at exactly $t = 0$, then one of them is chosen with probability $1/m$ to receive the payoff λ,and the other $m - 1$ players receive $\delta \times \lambda$. If no player stops the clock (and the game terminates at time $t = T$), then the payoff for each of the n players is g, where $0 \leq g < [\lambda \times (2^{(T/\theta)})]$. The parameters λ and θ control the magnitude and rate of increase in the payoff function, and g controls the incentive to let the clock run to time T. The parameters n, T, θ, λ, δ, and g are all commonly known, as is the form of the payoff function.

An example RTTG

Consider the RTTG with parameter values $T = 45$ (measured in seconds), $\theta = 5$, $\lambda = 5$, $\delta = 0.1$, and $g = 0$. Thus, each of the $n - 1$ "losers" receives 10% of the winner's payoff, and the payoff is 0 for all the n players if the clock is not stopped before 45 seconds. Payoffs are in cents. For any value of $n \geq 2$, if one of the players (say, 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) are:

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

[1] 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.

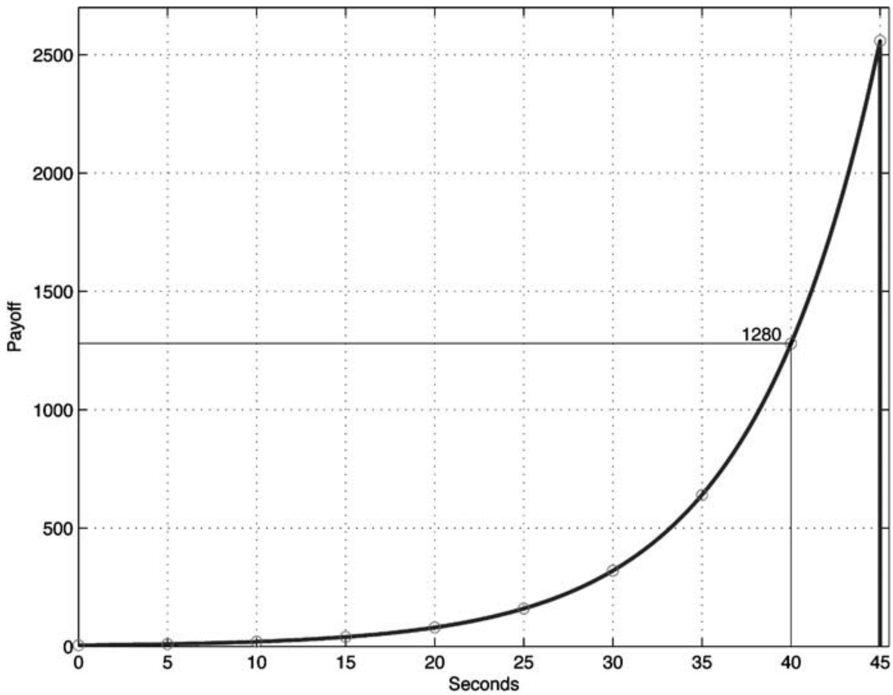


Fig. 2 Real-time trust game (RTTG) payoff function
Time (horizontal axis) is measured in seconds and the winner's payoff (vertical axis) is in cents. Consistent with the example, $t = 40$ *is marked and the corresponding payoff of 1280 is shown.*

If no player stops the clock, then each player receives $g = 0$. If m players stop the clock at exactly $t = 0$, then one of them is chosen with probability $1/m$ to receive the payoff of $\lambda = 5$. Thus, a winner can earn between \$0.05, if she stops the clock at exactly $t = 0$, and almost \$25.60, if she stops the clock just before 45 seconds.

Figure 2 exhibits the payoff function for this example; it is identical to the display subjects observed during actual play of the game. 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 winner in this particular example stopped the clock at 40.00 seconds and received the payoff $5 \times (2^{(40/5)}) = \$12.80$. Each of the two losers ($n = 3$ in this case) received \$1.28. If the clock were to be stopped at $t = 20$ seconds, then the winner would have received just \$0.80, and each loser only \$0.08. This instantiation of the RTTG has a payoff structure that is isomorphic to that in the three-player centipede game used by Rapoport et al. (2003).

As the RTTG is a new institution that was used previously only in the context of experiments on duels, Dutch auctions, and real-time public good games, our purpose is to test the effects of two of its parameters on behavior. We hypothesize that the population level of cooperation in this class of trust dilemmas will decrease in the group size, n, and increase in the value of the loser's fraction of the payoff, δ.

Equilibrium analysis of the RTTG

Let $r(t)$ denote the payoff to the winner who stops at time t ($t \geq 0$), and let $\delta \times r(t)$ denote the payoff to each of the remaining $n - 1$ players (losers). The payoff function $r(t)$ is strictly increasing, continuous, and uniquely defined for all $t \geq 0$. Because for small ε, $r(t)$ is greater than $\delta \times r(t + \varepsilon)$ for any value of t, it behooves each player to stop the clock before any of her opponents. This implies that in equilibrium every player should stop at time $t = 0$.

A RTTG experiment

Method

Subjects

One hundred and twenty-six subjects participated in six sessions, each including 21 subjects. Male and female students participated in about equal proportions. The subjects were recruited from undergraduate classes of business administration. Of the 126 subjects, 28 were offered partial class credit for showing up to the experiment *on time*. Prior to the session, the subjects were given the option of leaving the laboratory (without penalty) after receiving their non-salient payment. With a few exceptions, the subjects who arrived at the experiment opted to remain and take part in it for monetary reward. Each subject participated in a single session only, and all the subjects were granted anonymity. Payoff was contingent on performance. Including the show-up payment of \$5.00, individual earnings varied from \$7 to \$34.

Procedure

All six sessions were conducted at the Economic Science Laboratory at the University of Arizona, which contains 40 networked PCs in individual cubicles. At the beginning of each session, the subjects drew poker chips from a bag containing chips numbered from 1 to 21 to randomly determine their seat assignment in the laboratory. Each cubicle contained a PC and written instructions (see Appendix 1). Once seated, the subjects proceeded to read the instructions at their own pace. When they completed reading the experiment started.

Three experimental conditions were conducted that differed from one another in the values of n or δ; all other parameters of the RTTG remained the same.

Condition $n3/\delta0.5$:	$n = 3$ and $\delta = 0.5$	Larger δ condition
Condition $n3/\delta0.1$:	$n = 3$ and $\delta = 0.1$	Baseline condition
Condition $n7/\delta0.1$:	$n = 7$ and $\delta = 0.1$	Larger n condition

Each condition was replicated twice for a total of six sessions. Each session consisted of 90 rounds, except of the first session of Condition $n7/\delta0.1$ which was terminated by the experimenter after 47 rounds.[2] The parameter values common to all three conditions were $T = 45$ seconds, $\theta = 5$, $\lambda = 5$ and $g = 0$ (all the same as in the example presented before).

A random assignment procedure was used to determine group membership over iterations of the game. On each round, the 21 population members of a session were randomly assigned

[2] The reason for its early termination is discussed in the Results section below.

to seven groups of three members each (Conditions $n3/\delta0.1$ and $n3/\delta0.5$) or into three groups of seven members each (Condition $n7/\delta0.1$). At the beginning of each round, the subjects were only informed of the round number. The subjects were not allowed to communicate with one another. They were never informed of the identity of their group members. The random matching procedure was common knowledge; it was used to prevent reputation building. This environment does not support direct reciprocity between players, but players could deduce that they would likely be matched with each other repeatedly. As such, they might have been enticed to play cooperatively, thus fostering indirect reciprocity that would operate at the population level.

Given that elapsed time is a critical aspect of the RTTG, extra attention was given to the initiation phase of the experiment to ensure all subject were alert at the precise moment in which each trial began. To accomplish this objective, once all the 21 subjects indicated their readiness to start the game (by pressing a "Ready" button that appeared on the screen), a solid aqua-colored screen began flashing on the monitor as a warning sign that the round was about to begin. Several seconds later, the main screen was displayed on the monitor with indicator lights[3] bordering the sides of the main screen. Just prior to the clock (and payoff pot) starting, the red lights flashed four times, followed by a rapid series of four yellow lights, and then a solid green light. This warning phase lasted about three seconds. As soon as the green light appeared, the clock was started, and simultaneously the computer moved the mouse pointer to the center of a small rectangular white box placed in the middle of a larger red circle (see Appendix 2 for an image of what the subjects observed). To stop the clock, the subject only had to move the mouse pointer outside the white box. We used this procedure to eliminate any noise associated with clicking the mouse that might have conveyed signals about stopping times to other players in the laboratory. To follow the advance of the clock, a subject could attend either to the red line of the graph that increased exponentially, the clock box (that was accurate up to 1/20 of a second), or the payoff pot (with the same level of accuracy as the clock). We chose to fix T at 45 seconds because the pace was judged to be not too fast to impose time pressure yet not too slow to induce boredom.

We hypothesized that increasing δ would be conducive to establishing trust based cooperation among the players. As δ approaches 1, the opportunity cost of misplaced trust decreases, making the exercise of trust less risky. Conversely we anticipated that increasing n would be detrimental to the development of trust. Trust interactions have a weakest link property, and groups of larger size are clearly more vulnerable to defection of one of its members than groups of smaller size. This hypothesis is consistent with results from weak-link games (Weber et al. 2001) and the findings of Bonacich et al. (1976) in their study of n-player prisoner dilemma games.

Results

Aggregate level analysis

The distribution of stopping times in Conditions $n3/\delta0.1$ was compared to the distribution of "exit moves" in a centipede game from the second experiment of Rapoport et al. (2003). The purpose of this comparison was to establish that trust-based behavior is roughly the same in a discrete vs. continuous strategy environment, ceteris paribus. Recall that the baseline

[3] Modeled after the warning lights commonly used to prepare drivers for a prompt start at professional drag racing competitions.

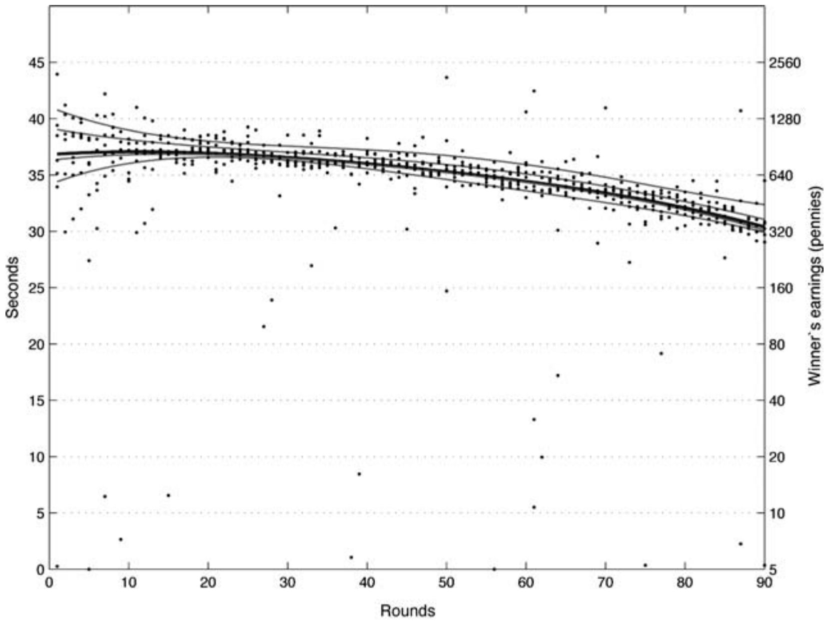
condition of the RTTG used here has an isomorphic payoff structure to the three-player centipede game. In order to compare the two experiments, data from the RTTG were discretized into bins of five-second width. This resulted in nine bins, comparable to the endnodes of the 3-player centipede game. The relative proportions of exit decisions were then compared between the two experiments using a Kolmogorov-Smirnov two-sample test. The test yielded non-significant results, indicating that the distribution of ending moves (exit moves or stopping times) across experiments were not significantly different.

We recorded the stopping times (seven for each round in Conditions $n3/\delta0.1$ and $n3/\delta0.5$, and three in Condition $n7/\delta0.1$) that could assume any value between $t = 0$ and $t = 45 - \varepsilon$ seconds. Altogether, there were $7 \times 90 = 630$ data points for each session in Conditions $n3/\delta0.1$ and $n3/\delta0.5$. The first session of Condition $n7/\delta0.1$ had $3 \times 47 = 141$ data points, and the second session of Condition $n7/\delta0.1$ had $3 \times 90 = 270$ data points. Figure 3 displays all the data points for the first and second sessions of Condition $n3/\delta0.5$. The raw data for the first and second sessions are shown in the top and bottom parts of Fig. 3, respectively. To display the trends in stopping times across rounds, 4th order polynomial functions[4] were fitted separately for the 2nd, 3rd, 4th (median), 5th, and 6th ranked stopping times using ordinary least squares. The lowest and highest stopping times were not fitted because of their considerable variability. Figure 3 shows that stopping times between 0 and 5 seconds occurred periodically; we attribute them to error, impatience, or sheer frustration.

Subjects in both sessions of Condition $n3/\delta0.5$ behaved in a similar manner. Stopping times of the seven groups were quite variable during the first 10–15 rounds, and then they stabilized. Two major patterns emerged quite clearly. First, the average stopping time decreased slowly across rounds from about 35 to 30 seconds. Small as it may appear, this five-second difference translates to a 50% drop in the winner's average payoff from \$6.40 to \$3.20. The decrease in median stopping time is seen to be very steady, with the difference in the median stopping time between two consecutive rounds measured in fractions of a second. The subjects clearly realized that it was to their mutual benefit to stop the clock as late as possible. However, the temptation to stop and win the prize at time t, rather than wait and get 50% of a higher but otherwise unknown prize at time $t + \varepsilon$, was simply too strong for many subjects, and it overcame the desire to let the clock run. We interpret this finding as a gradual breakdown of trust-based cooperation in a population of anonymous players divided into small groups where neither reputation building nor direct punishment for breaking trust is possible. The existence of a few "hard core" (i.e. dogmatic) cooperators can be gleaned in Session 1, where in seven cases the clock was stopped after about 43 seconds (with a considerable increase in payoff). When three such players happened to be randomly assigned to the same group, then occasionally they let the clock run and then stopped it at a time that came close to maximizing joint payoffs.

The second, and for us quite unexpected, finding were the small differences in stopping times among the seven groups after about 15 rounds. Recall that (1) subjects were randomly assigned to 3-member groups at the beginning of each round; and, (2) they were only informed of the stopping time of their own group after each round. Figure 3 shows that, with only few exceptions, stopping times after round 15 fell into an interval that seldom exceeded three seconds. Because of the intermixing of the members of each population, a strong "social norm" was established in the population dictating the time to stop the clock (or, equivalently, the winner's payoff). Although average stopping time slowly decreased over iterations of the

[4] The order of the polynomial was chosen *ad hoc*. On visual inspection, 4th order polynomials seemed to fit the data sufficiently while not being overly susceptible to random perturbations. Fourth-order polynomials are used throughout the paper.



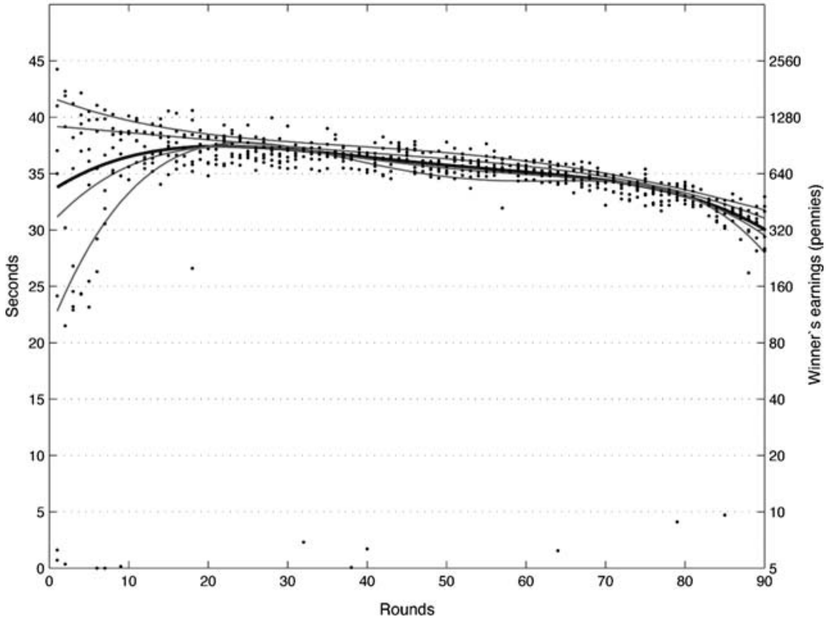


Fig. 3 Stopping times by round in condition $n3/\delta0.5$, sessions 1 and 2
Stopping times (linear scale on the left y-axis) and corresponding payoffs to the winner of each group (logarithmic scale on the right y-axis) from Condition n3/δ0.5. The experiment includes 90 rounds and 7 groups of $n = 3$ players each. Therefore, each round contains 7 stopping times. Forth order polynomial trend lines are fitted to the data to highlight the dynamics of the experimental population.

stage game, the variability remained more or less constant. (The slight fanning out at the last 3–4 rounds in Session 2 is an end effect, as it was commonly known that the experiment consisted of 90 rounds.)

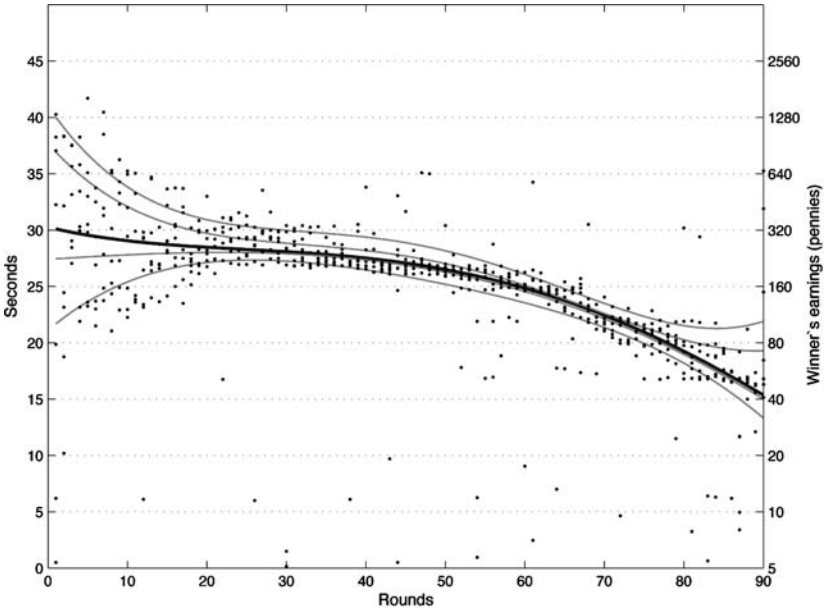
Using the same format as Fig. 3, Fig. 4 exhibits all the stopping times–a total of 630 data points for each session–for the two sessions of Condition $n3/\delta0.1$. The same two general patterns observed in Fig. 3 are also discernable in Fig. 4. First, average stopping time steadily decreased over rounds. Second, after about 20 rounds, stopping times of all seven groups became very close to one another. There are three minor differences and one major difference between Conditions $n3/\delta$ 0.1 and $n3/\delta0.5$. First, the establishment of a "stopping norm" required a few more rounds in Condition $n3/\delta0.1$ than in Condition $n3/\delta0.5$. Second, the fanning out of the stopping times toward the end of the session was more pronounced in Condition $n3/\delta0.1$ than in Condition $n3/\delta0.5$ and started earlier. Third, we observe a higher frequency of stopping times below 10 seconds in Condition $n3/\delta0.1$ than in Condition $n3/\delta0.5$.

The major difference between the two conditions is that the clock was stopped in Condition $n3/\delta0.1$ significantly earlier than in Condition $n3/\delta0.5$. Median stopping time started at about 31 seconds and then dropped to 15 seconds in each of the two sessions. This change in median stopping time from round 1 to round 90 translates to a considerable reduction in average payoff–four times as large as the decay in Condition $n3/\delta0.5$–from about \$3.40 to a meager \$0.40.

To formally compare the two conditions to each other, we computed the median stopping time among the seven triads for each round ($n = 90$ observations in each session). Using the non-parametric Mann-Whitney test[5], the null hypothesis of equality between the median stopping times was rejected ($z = -15.35$, $p <0.001$). Reducing the loser's fraction of the winner's payoff from 50% to 10% resulted in significantly lower stopping times and, consequently, lower payoffs for all subjects. Toward the end of the session, subjects in Condition $n3/\delta0.1$ were earning 1/8th of the subjects in Condition $n3/\delta0.5$. Moreover, the decrease in the median payoffs for the winner accelerated faster when the loser's share was reduced from 50% to 10%.

Figure 5 displays the raw data for the two sessions of Condition $n7/\delta0.1$. Because only three data points are recorded for each round, a 4th order polynomial was fitted only to the median stopping times. Increasing group size from 3 to 7, while keeping the loser's share of the payoff at 10% (the same as in Condition $n3/\delta0.1$), resulted in dramatic differences between the two conditions. In both sessions of Condition $n7/\delta0.1$, the median stopping times started at 17 seconds (about \$0.53) and then declined to zero. The two sessions only differed from each other in the rate of decrease in stopping times. In Session 1, the stopping times unraveled rapidly reaching equilibrium at $t = 0$ in less than 30 rounds. Once it became evidently clear that no recovery was likely (and after observing "clear signs of irritation" in some of the subjects), the experimenter terminated the session after 47 rounds. In contrast, it took about 70 rounds until the stopping times in Session 2 converged to zero. In this case (absent obvious disgruntled subject behavior), the session ran through completion (90 rounds as initially stated in the instructions). The null hypothesis of equality in median stopping times of Conditions $n3/\delta0.1$ and $n7/\delta0.1$ was rejected ($z = 14.13$, $p <0.001$). The conclusion drawn from this comparison is that trust-based cooperation, when reputations cannot be built, strongly depends on group size and deteriorates faster as n increases.

[5] The assumption of independence is not clearly met by these data so the results of these significance tests should be interpreted with caution. However, the results from the tests indicate large effects (i.e. there is not a problem with marginal significance) and the results corroborate conclusions drawn from visual inspection of the data.



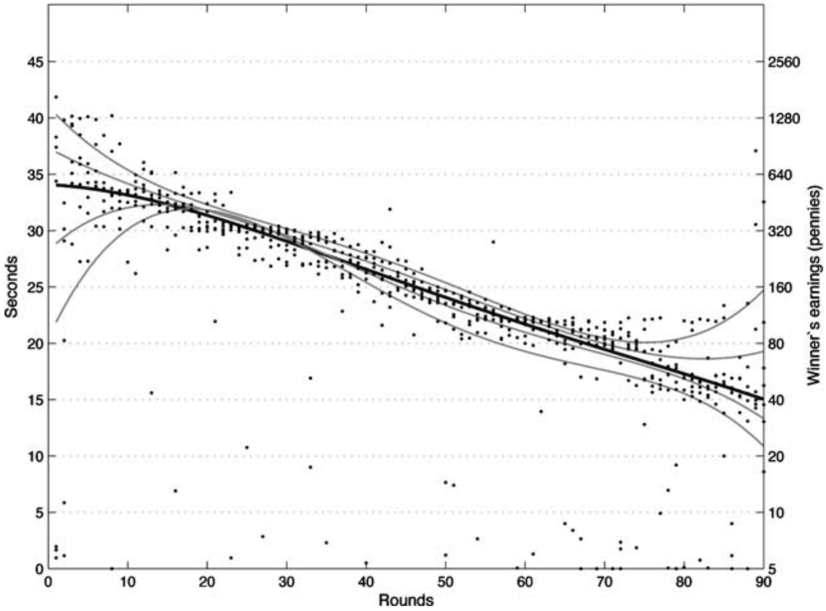
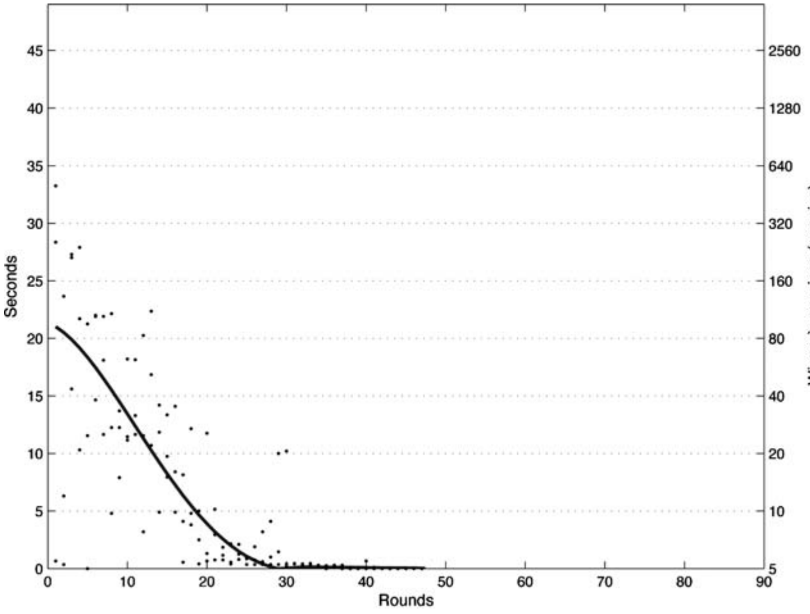


Fig. 4 Stopping times by round in condition $n3/\delta0.1$, sessions 1 and 2
Stopping times (linear scale on the left y-axis) and corresponding payoffs to the winner of each group (logarithmic scale on the right y-axis) from Condition n3/δ0.1. The experiment includes 90 rounds and 7 groups of $n = 3$ players each. Therefore, each round contains 7 stopping times. Forth order polynomial trend lines are fitted to the data to highlight the dynamics of the experimental population.



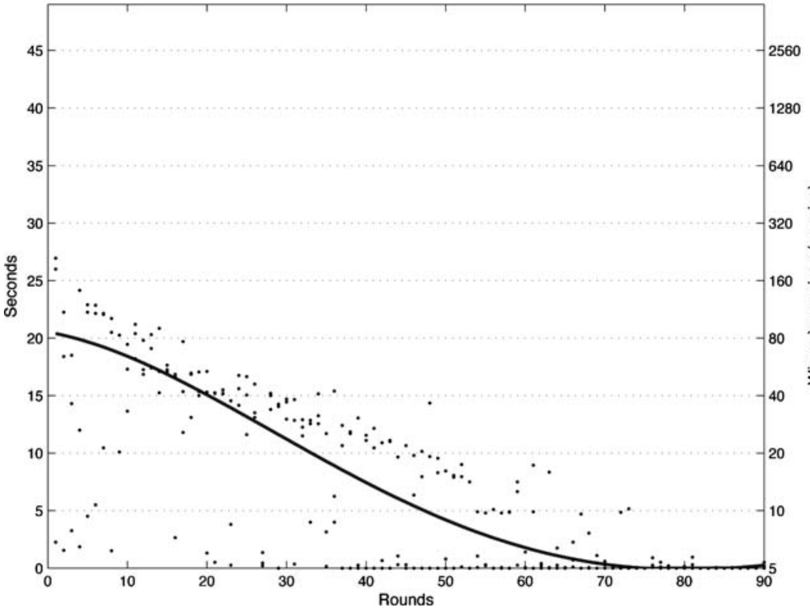


Fig. 5 Stopping times by round in condition $n7/\delta0.1$, sessions 1 and 2
Stopping times (linear scale on the left y-axis)and corresponding payoffs to the winner of each group (logarithmic scale on the right y-axis) from Condition n3/δ0.5. The experiment includes 47 and 90 rounds respectively with 3 groups of n = 7 players each. Therefore, each round contains only 3 stopping times. A single forth order polynomial trend line is fitted to the median stopping time per round in an effort to highlight the dynamics of the experimental population.

Three additional Mann-Whitney tests were conducted to determine if there were statistically significant differences between sessions within condition. All three tests (one for each experimental condition) yielded non-significant results.

Individual level analysis

The individual frequency of stopping the clock could take any value between 0 and 90: 0, if the subject never stopped the clock; 90, if she did so on each of the 90 rounds of the experiment. The narrow bands of stopping times in both Conditions $n3/\delta0.1$ and $n3/\delta0.5$, which seldom exceeded 3 seconds, would seem to suggest only minor differences among the subjects in the frequencies of stopping times. In fact, this was not the case at all, as some subjects showed an uncanny ability to guess the intentions of their group members (who were randomly assigned to groups on each round) and preceded them in being the first to stop the clock by a fraction of a second. The individual frequencies of stopping the clock varied between 1 and 62 in Session 1 and between 9 and 61 in Session 2 of Condition $n3/\delta0.5$. The corresponding frequencies for Sessions 1 and 2 of Conditions $n3/\delta0.1$ were 6 to 52 and 9 to 57.

Figure 6 shows the cumulative frequency distributions of individual frequencies of stopping the clock. Two cumulative distributions are displayed, one across both sessions of Condition $n3/\delta0.5$ and the other across both sessions of Condition $n3/\delta0.1$. The horizontal

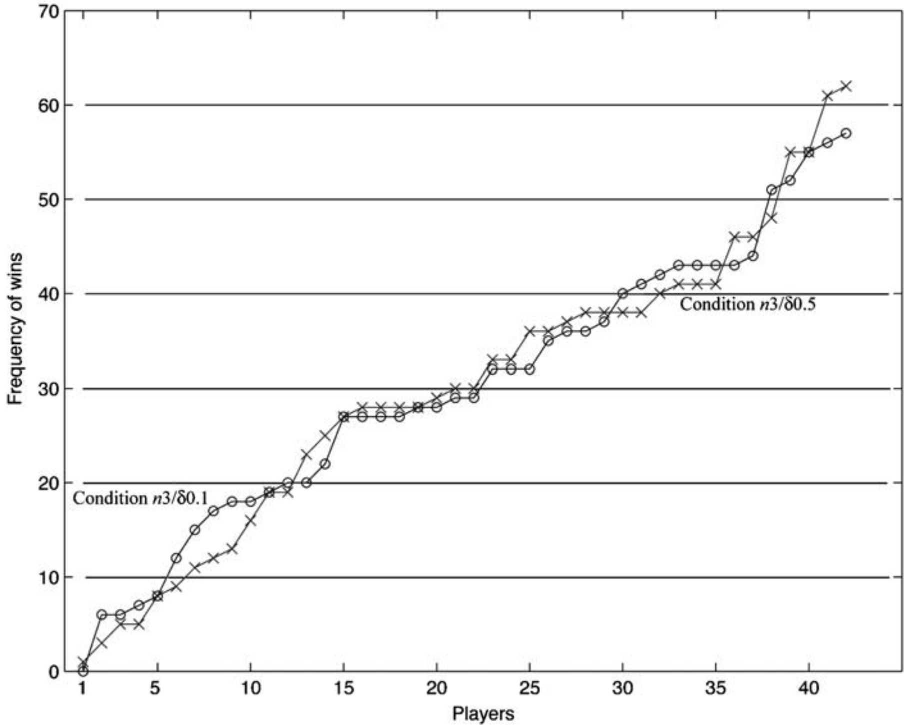


Fig. 6 Cumulative frequency distributions of subject's wins across conditions $n3/\delta0.1$ and $n3/\delta0.5$
The x-axis corresponds to the rank ordering of subjects within each condition according to the number of wins for each. The y-axis shows the frequency of wins. As can be seen, there is a single subject in Condition n3/δ0.1 who never won (i.e. never stopped the clock). On the other extreme, there is a subject in Condition n3/δ0.5 who won on 62 of the 90 rounds. The two conditions are not significantly different from each other.

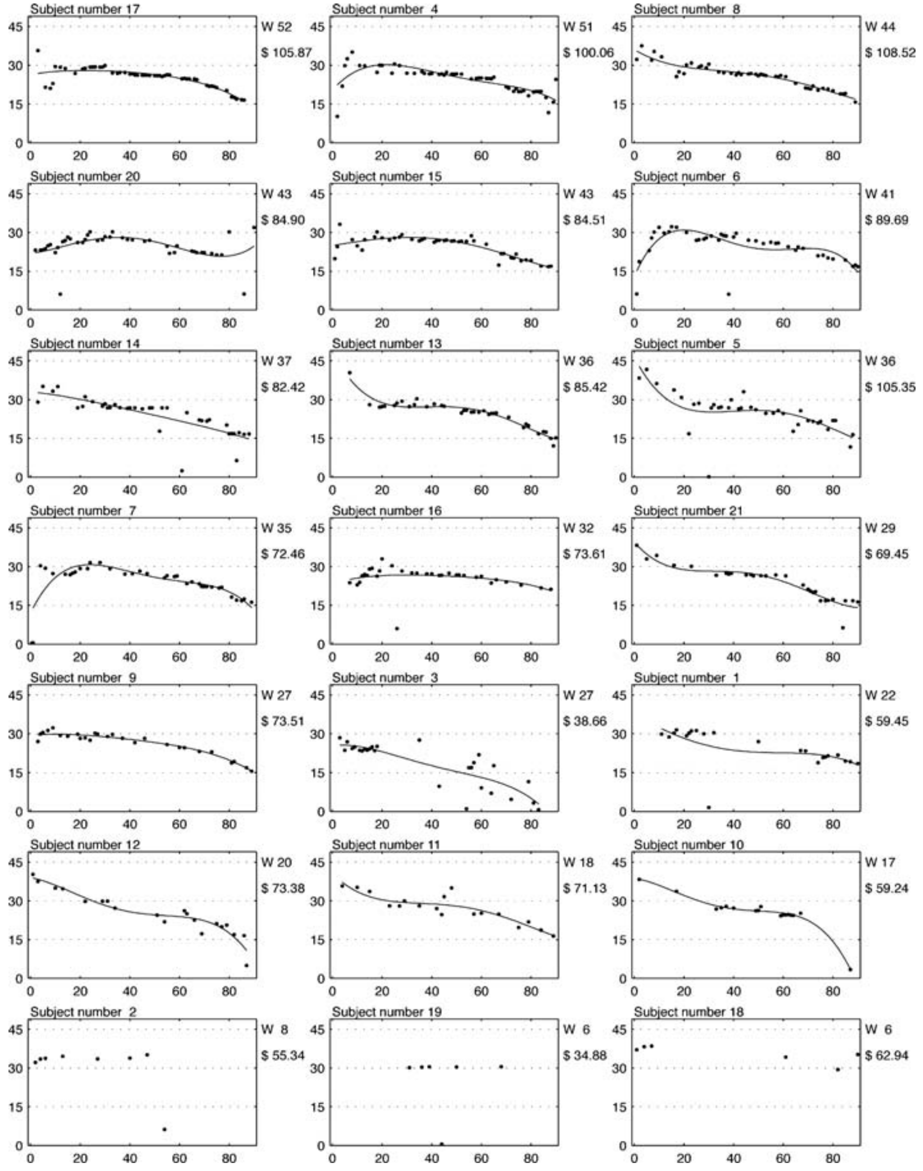


Fig. 7 Individual plots of stopping times in condition $n3/\delta$ 0.1, session 1

axis marks the rank of the subject in each condition from 1 to 42 (2 × 21) in terms of the number of wins; a lower rank indicates fewer wins. The vertical axis shows the individual frequencies of stopping the clock across all rounds (range: 0 to 90). The null hypothesis that the two cumulative frequency distributions do not differ from each other could not be rejected by a two-sample Kolmogorov-Smirnov test ($D = 0.436$, $p = 0.991$). This finding suggests no differences between the two samples of subjects in terms of individual propensities to cooperate as manifested by waiting to stop the clock. The only difference between the two conditions (compare Figs. 3 and 4) is in *how long* subjects were willing to trust their group members and, rather than stop the clock, wait for larger payoffs as the loser's fraction of the payoff, δ, increases.

Figure 7 exhibits individual plots of stopping times for the 21 subjects in Session 1 of Condition $n3/\delta0.1$. Individual plots for the other five sessions do not provide any new information and are, therefore, not displayed. The same 4th order polynomial function was fitted to the individual stopping times for subjects who stopped the clock at least 15 times. A comparison of the individual plots in Fig. 7 with the plot at the top part of Fig. 4 shows that most of the individual functions are very similar in shape to the function depicting the median stopping time, starting at about 30 seconds on round 1 and ending at about 15 seconds at round 90. Most of the exceptions concern subjects who stopped the clock less than 20 times.

Each individual plot in Fig. 7 also displays the total frequency of stopping (labeled *W*for wins) per subject, and the total payoff that any subject would have earned if all 90 rounds (rather than a random sample of six rounds) were counted for payment. The hypothetical individual payoffs thus computed range from \$34.88 (Subject 19) to \$108.52 (Subject 3). As expected, the total payoff across all 90 rounds is positively and significantly correlated with the frequency of stopping the clock ($r = +0.85$ for results combined across both sessions of Condition $n3/\delta0.1$).

Conclusions

The RTTG has been devised to study the dynamics of trust-based cooperation in a population of agents participating in trust dilemmas that share three major features. First, as long as all the n players in a group continue their cooperation their joint payoff increases exponentially over time. Second, the temptation to defect ("exit") increases at the same exponential rate, with the first player to defect receiving the lion's share of the joint payoff and each of the other $n - 1$ players only receiving a fraction of that player's payoff. Third, as the game ends with a single stopping decision, neither punishment nor direct reciprocity are possible. In equilibrium, each player should exit at time $t = 0$. Unlike the centipede game and related extensive-form games that have been proposed to study trust and trustworthiness, players are no longer treated asymmetrically. Consequently, each player has the same opportunity to contribute to the buildup of trust-based cooperation or cause its breakdown. If δ–the fraction allotted to the loser–is set at zero, then the RTTG resembles the Dutch (descending) auction, where each player's payoff increases as the clock runs longer, each can stop the clock, and only the one who does so receives a non-zero payoff. If $n = 2$, then the RTTG is similar to a noisy duel (e.g., Kahan & Rapoport, 1974) with equal exponential accuracy functions. The RTTG also has shares common features several real-time public goods games developed by Kurzban et al. (2001) and subsequently studied by Goren et al. (2003), Goren et al. (2004) Ishii and Kurzban (2005).

Our results show that if $n = 3$, then on round 1 players start with a propensity to trust their group members by waiting for 30–35 seconds before stopping the clock with corresponding payoff to the winner ranging between \$3.20 and \$6.40. These results are similar to the results of the Berg et al. (1995) investment game, where the first movers are willing, on average, to trust the second movers with a substantial fraction of their endowment. Our results further show that populations are unlikely to maintain this initial level of trust-based cooperation with iterations of the stage game when group membership is changed randomly from round to round. Rather, cooperation *in the population* breaks down gradually, with the rate of breakdown increasing sharply in n and decreasing more slowly in δ. We observe attempts by a few hard core (dogmatic) cooperators to reverse this downside trend by delaying their exit decisions. Occasionally when these hard core cooperators happen to be assigned to the same group by the random matching design–an event that is more likely to happen when $n = 3$ than $n = 7$–they let the clock run longer and thereby increase their joint profit. But these attempts are, in general, not sufficient to overcome the greed associated with winning.

Caution should be exercised in generalizing these disheartening results beyond the experimental design and the parameter values of the present study. Additional experiments could be conducted for better understanding the effects of the variables that govern the evolvement and breakdown of trust-based cooperation in this class of trust dilemmas. We briefly mention a few. First, as already suggested earlier, the exponential payoff function could be replaced by and compared to some other monotonically increasing function, e.g., a linear payoff function. We have opted to use the exponential function primarily in order to replicate the payoff function used by Rapoport et al. (see Fig. 1). Second, as suggested by some of the people who read an earlier version of this paper, a stronger case for the claim that our study investigates trust could be made by setting g at a positive value. An important question is whether the downtrends exhibited in Figs. 3-5 for all six sessions could be stopped or even reversed if the payoff g for never defecting is set at a sufficiently high positive value. Third, the random matching design may be replaced by a fixed-group design. In this case, we would expect larger between-group differences in the stopping time than those observed in the present study as each group may develop its own norm. Finally, the *decision method* used in the present study, where only a single player has the opportunity to record her intended stopping time, may be replaced by the *strategy method* where each player is required to state her intended stopping time before the game begins at $t = 0$, and then the clock is run until time T.

Acknowledgments We gratefully acknowledge financial support of this research by a grant from the Hong Kong Research Grants Council (Project No. HKUST6225/99H) to the Hong Kong University of Science and Technology.

References

Arrow, K. (1974). *The Limits of Organizations*. NY: Norton.

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.* Cambridge, MA: MIT Press, pp. 214–227.

Aumann, R. J. (1995). Backward induction and common knowledge of rationality. *Games and Economic Behavior*, *8*, 6–19.

Aumann, R. J. (1998). On the centipede game. *Games and Economic Behavior*, *23*, 97–105.

Bacharach, M., Guerra, G., & Zizzo, D. J. (2001). Is trust self-fulfilling? An experimental study. Oxford University, Department of Economics, unpublished manuscript.

Ben-Porath, E. (1997). Rationality, Nash equilibrium and backwards induction in perfect-information games. *Review of Economic Studies*, *64,* 23–46.

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

Bonacich, P., Shure, G., Kahan, J., & Meeker, R., (1976). Cooperation and Group Size in the n-Person Prisoners' Dilemma. *Journal of Conflict Resolution*, *20*, 4, 687–706.

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

Camerer, C. F. (2003). *Behavioral Game Theory: Experiments in Strategic Interaction*. NY: Russell sage Foundation.

Camerer, C. F. & Weigelt, H. K. (1988). Experimental tests of a sequential equilibrium reputation model. *Econometrica*, *56*, 1–36.

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

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

Fey, M., McKelvey, R. D., & Palfrey, T. R. (1996). An experimental study of constant-sum centipede games. *International Journal of Game Theory*, *25*, 269–287.

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

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

Goren, H., Kurzban, R., & 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., & 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., & 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., & 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., & Wendel, M. (1997). Cooperation based on trust: An experimental investigation. *Journal of Economic Psychology*, *18*, 15–43.

Ho, T. H., & Weigelt, K. (2001). *Trust building among strangers*. University of Pennsylvania, unpublished manuscript.

Ishii, K., & Kurzban, R. O. (2005). Real time pubic good games in Japan: Cultural and individual differences in trust and reciprocity. University of Pennsylvania, Dept. of Psychology, unpublished manuscript.

James, H. S. (2002). The trust paradox: A survey of economic inquiries into the nature of trust and trustworthiness. *Journal of Economic Behavior and Organization*, *47*, 291–307.

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

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, & Y. H. Tan (eds.), *Trust in Cyber-Societies*. Berlin: Springer-Verlag, pp. 9–26.

Kurzban, R., McCabe, K. A., Smith, V. L., & Wilson, B. (2001). Incremental Commitment and Reciprocity in a Real-Time Public Goods Game. *Personality & Social Psychology Bulletin*, *27*, 12, 1662–1673.

Maskin (eds.), *Economic Analysis of Markets & Games: Essays in Honor of Frank Hahn*. Cambridge, MA: MIT Press, pp. 214–227.

McCabe, K. A., Rassenti, S. J., & 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., & 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., & Smith, V. L. (2002). Cooperation in single play, two-person extensive form games between anonymously matched players. In R. Zwick & A. Rapoport (eds.), *Experimental Business Research*. NY: Kluwer, pp. 51–67.

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

McKelvey, R., & Palfrey, T. (1992). An experimental study of the centipede game. *Econometrica*, *60*, 803–836.

Nagel, R., & Tang, F. F. (1998). Experimental results on the centipede game in normal form: An investigation of learning. *Journal of Mathematical Psychology*, *42*, 356–384.

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

Ponti, G. (2000). Cycles of learning in the centipede game. *Games and Economic Behavior*, *30*, 115–141.

Rapoport, A. (2003). Centipede games. In L. Nadel (ed.), *Encyclopedia of Cognitive Science, Vol. 2*. London: Macmillan, pp. 196–203.

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

Reny, P. J. (1993). Common belief & the theory of games with perfect information. *Journal of Economic Theory*, *59*, 257–274.

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., & Camerer, C. (1998). Not so different after all: A Cross-Discipline View of Trust, *Academy of Management Review*, *23*, 3, 393–404.

Rubinstein, A. (1982). Perfect equilibrium in a bargaining model. *Econometrica*, *50*, 97–109.

Stalnaker, R. (1996). Knowledge, belief, and counterfactual reasoning in games. *Economics and Philosophy*, *12,* 133–163.

Stalnaker, R. (1998). Belief revision in games: Forward and backward induction. *Mathematical Social Sciences*, *36*, 31–56.

Weber, R., Camerer, C., Rottenstreich, Y., & Knez, M. (2001). The Illusion of Leadership: Misattribution of Cause in Coordination Games. *Organization Science*, *12*, 5, 582–598.