In a long sequence of iterated Prisoner's Dilemma played by a pair of subjects, "lock-in" effects are frequently observed both on the CC and on the DD outcome. That is, in the course of the sequence either CC or DD becomes predominant sometimes persisting to the end. A lock-in on CC can be explained by a deterrence effect: The subjects learn that defection to D, although immediately advantageous if the co-player continues with C, is likely to be followed by retaliation. A lock-in on DD can be explained by the inhibition of unilateral C (the "peace initiative"), which is associated with the sucker's payoff. In the present experiment, subjects played repeated round robin sequences of Prisoner's Dilemma with other subjects in their experimental group. Prominent lock-in effects were observed when the subjects were informed on each play how their current co-players chose when paired with other subjects but not when no such information was given.

Experiments With Social Traps IV

REPUTATION EFFECTS IN THE EVOLUTION OF COOPERATION

ANATOL RAPOPORT
University of Toronto
ANDREAS DIEKMANN
AXEL FRANZEN
University of Bern, Switzerland

In his discussion of the evolution of cooperation, Axelrod (1984) emphasized the importance of repeated interactions between the same individuals. In particular, he cites examples of tacit truces between soldiers in trenches on the western front during World War I, who developed live and let live policies by refraining from shooting. The high command did not approve these peace initiatives and tried to prevent them by frequent rotation of units, so that soldiers would face enemies rather than people whom they had come to regard as colleagues who shared their misfortunes.

Similar effects have been observed in iterated plays of Prisoner's Dilemma. In long series of such plays, subjects frequently lock in on the CC outcome. Although this outcome is not an equilibrium of the game iterated a

RATIONALITY AND SOCIETY, Vol. 7 No. 4, October 1995 431-441

known finite number of times, it is presumably preserved by a deterrence effect. The temptation to defect to D is resisted by the realization that the defection is very likely to be retaliated, resulting in DD disadvantageous to both players. Lock-ins are consistent with Axelrod's hypothesis that repeated encounters between the same individuals enhance opportunities for establishing cooperation. To be sure, lock-ins on DD are also frequently observed. Here, players are deterred from attempts to initiate cooperation because such unilateral "peace initiatives" result in the "sucker's payoff" and, besides, reward the player who continues to defect. For humans, information may be an effective substitute for experience. In particular, if one has a choice of trusting or not trusting an actor with whom one is to interact, the latter's reputation may be a decisive factor. Moreover, if on the basis of such reputation, individuals are likely to cooperate, the tendency to cooperate further is enhanced in consequence of a positive feedback effect. Of course, the same holds for the tendency to withhold cooperation.

In economic models of human behavior, resort to cooperative strategies in the so-called social trap situations, exemplified by finitely iterated Prisoner's Dilemma, appears at first thought as a disconfirmation of a fundamental assumption underlying these models. Several authors have offered explanations of such behavior, based on reputation effects. Coleman (1988) called attention to the closed community of Jewish diamond merchants in New York, closely knit by intermarriage and religious affiliation, in which any violation of trust (in pursuit of large short-term gain) would result in immediate expulsion. Granovetter (1985) emphasized "embeddedness," the effect of social structures and networks on the behavior of individuals embedded in them. In particular, the role of reputation effects, aside from "socialization," usually invoked against assumptions underlying economic models, provided a way of bringing cooperative behavior in social traps under the umbrella of rational choice models.

Raub and Weesie (1990) extended these ideas to formal game-theoretic analysis. By introducing reputation effects generated by embeddedness of individuals in social networks, they showed that lock-ins on CC in iterated Prisoner's Dilemma appeared as an equilibrium in the resulting game. An important point in their discussion is the double role of the cooperative choice in a network of players informed of each other's actions: (a) as a signal to others that one can be trusted to choose C and (b) as evidence that the current co-player can be so trusted. The homily "Honesty is the best policy" is evidence of early recognition of the role played by reputation in business.

Policies based on building up one's image are not all designed to enhance cooperation. Their aim may also be to intimidate, for example, by convincing others of one's ruthlessness regardless of costs incurred, as in the game of

Chicken or in the Chain Store Paradox (Selten 1978). In this article, we shall be concerned with testing a specific consequence of the model developed by Raub and Weesie.

In the experiment to be described, performed partly in Toronto, partly in Bern, reputation effects in iterated Prisoner's Dilemma were assessed. Groups from 5 to 8 subjects (students of the University of Toronto or the University of Bern) participated in each session. The players sat in front of consoles, facing individual screens on which information on the course of the game was displayed. On each play of the game, each subject was paired with another subject in the group. In a round of the game, each subject played once against every other subject. Twenty-two rounds were played at every session.

The iterated game was played under two conditions. Under the FEEDBACK condition, on every play after the first, each subject was informed about his or her current co-player's choice on the previous 1 to 10 rounds. Under the NO FEEDBACK condition, this information was not given. The following rules, displayed on successive frames, describe the procedure.[1]

FRAME 1

You will be playing a game with other persons, your co-players. The game will last several rounds.[2] On each round, your co-player may be different. The players are numbered 1, 2, 3, and so on. You are player (number).

The game will be represented by the following diagram:

	S	T
S	(S, S)	(S, T)
T	(T, S)	(T, T)

On each, round you will choose either the top row (S) or the bottom row (T).

Simultaneously, your co-player will choose either the left-hand column (S) or the right-hand column (T). Your choice and your co-player's choice will determine one of the four outcomes of that round—one of the four boxes in the diagram—upper left (SS), upper right (ST), lower left (TS), or lower Tight (TT).

FRAME 2

In each box, two numbers will be shown. The first represents the number of points you have earned in that particular outcome. The second number represents the number of points earned by your co-player.

	S	T
S	300, 300	0, 500
T	500, 0	100, 100

For instance, if the outcome is SS, you and your co-player have each earned 300 points. If ST is the outcome, you have earned 0 points and your co-player has earned 500.

FRAME 3

	S	T
S	300, 300	0, 500
T	500, 0	100, 100

Let us see whether you have understood the rules. Suppose you chose S and your co-player chose T. How many points have you earned? Answer by depressing a number key on your console. (If the answer is 0, CORRECT! is displayed on the frame. Otherwise, WRONG, TRY AGAIN!)

After a number of such checks and a test run, the play begins. After each play, the following information is displayed:

FRAME 4

You chose. . . . Your co-player chose. . . .
You have earned. . . . Your co-player has earned. . . .

Under the FEEDBACK condition, preceding each play after the first, the players are given the following information:

FRAME 5

On round . . . your co-player chose. . . .
On round . . . your co-player chose. . . .
. . . (up to ten previous rounds).

Under the NO FEEDBACK CONDITION, this information was omitted. The points earned by each subject were converted to money at the rate of $1 Canadian per 3,000 points in Toronto, 1 Swiss Frank per 2,500 points in Bern. For showing up, subjects were paid $7 Canadian in Toronto, SF 10 in Bern.

Sixteen groups were run under the F condition and sixteen under the N condition. The principal hypothesis to be tested was that larger proportions of CC and DD outcomes would be observed under the F condition than under the N condition. Another aim of the experiment was to see whether some evidence of "evolution of cooperation" could be discerned and attributed to the FEEDBACK condition, that is, to the reputation effect.

RESULTS

Comparisons of the frequencies of CC and DD outcomes under the two conditions are shown in Figure 1. The principal hypothesis appears to be corroborated with respect to CC frequencies, but not with respect to DD frequencies. Evidently, the reason for the latter result is the overall increase of cooperation (reflected in the C frequencies and hence a decrease of D frequencies), as can be seen in Figure 2.

The overall effect of feedback on lock-ins (both CC and DD) is reflected in the binary correlations coefficient, which in the context of Prisoner's Dilemma is given by

$$r = \frac{(CC)(DD) - (CD)(DC)}{[(CC + CD)(CC + DC)(DD + CD)(DD + DC)]^{1/2}},$$

where (C), (D), (CC), and so on are the relative frequencies of the corresponding outcomes.

We note that when CD = DC = 0, that is, when both players lock in on either CC or DD, r = 1. On the other hand, when CC = (C)(C), CD = (C)(D), and so on, namely, when the choices of co-players are independent of each other, r = 0.

From Figure 3, we see that r is significantly greater than 0 under the FEEDBACK condition but very nearly 0 under the NO FEEDBACK condition, corroborating the reputation effect on lock-ins.

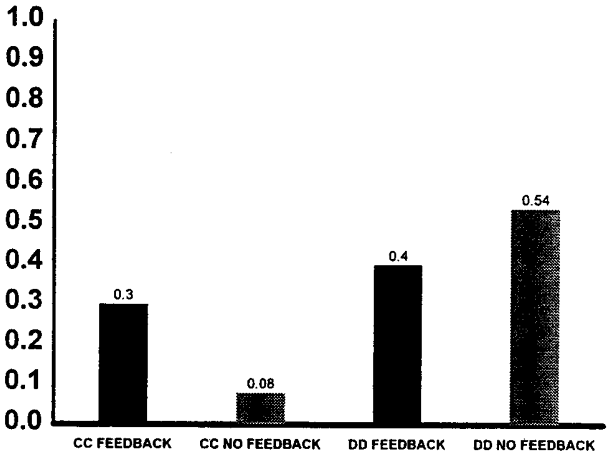

Figure 1: **Relative frequencies of CC and DD outcomes under FEEDBACK and NO FEEDBACK conditions (all data).**

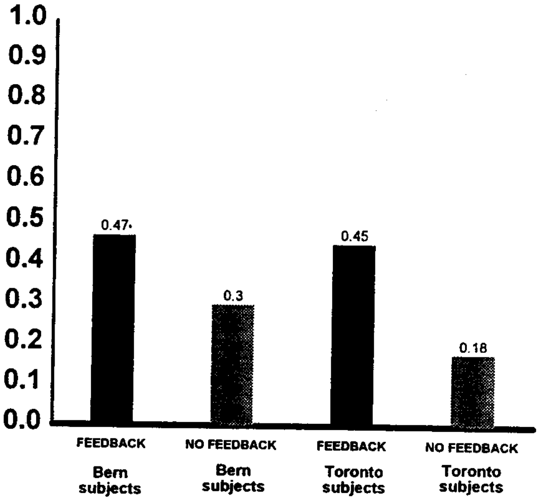

Figure 2: **Relative frequencies of C choices under FEEDBACK and NO FEEDBACK conditions.**

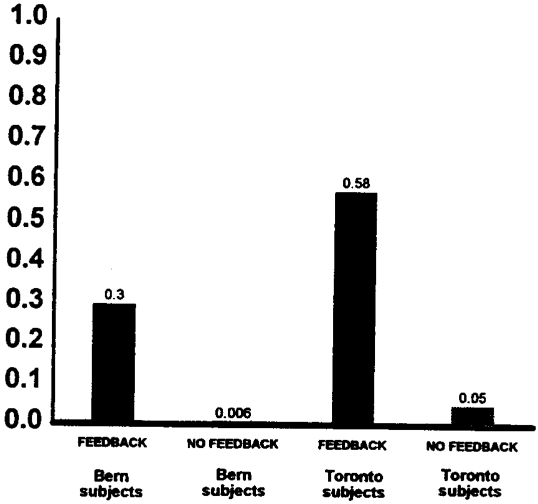

Figure 3: Comparison of binary correlation coefficients between C and D choices under FEEDBACK and NO FEEDBACK conditions.

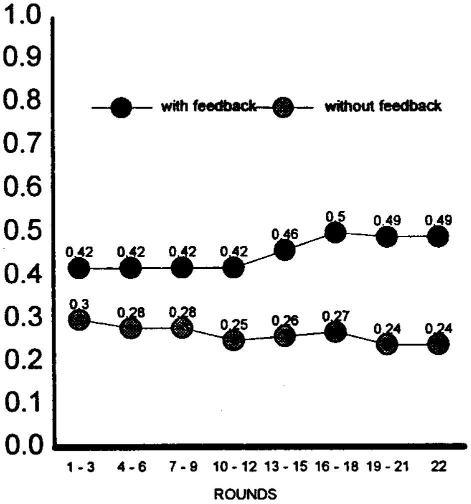

Figure 4: Time courses of C-choice frequencies (all data).

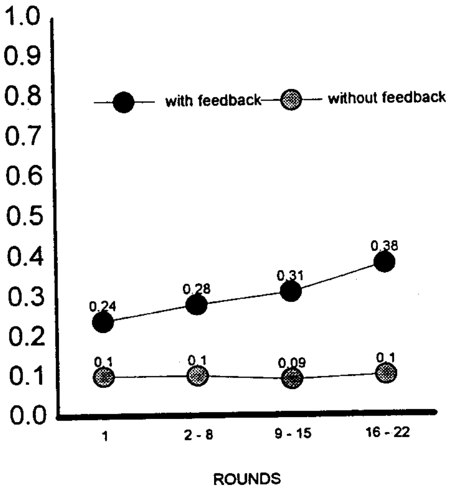

Figure 5: Time courses of CC outcome frequencies in seven 7-person groups.

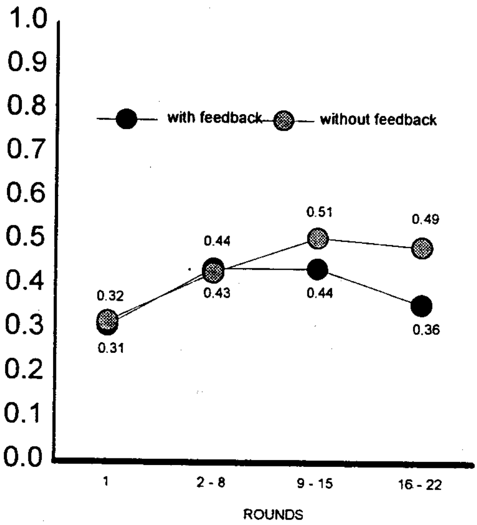

Figure 6: Time courses of DD outcome frequencies in eight 7-person groups.

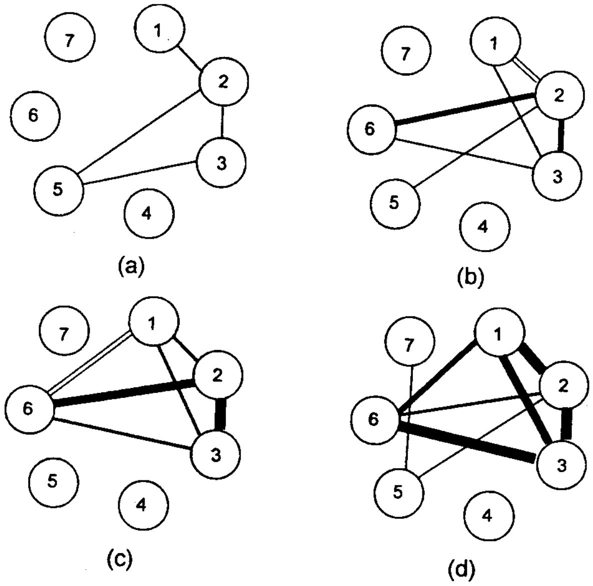

Figure 7: Cooperative links in a group playing under the FEEDBACK condition in successive stages: (a) round 1, (b) rounds 2-8, (c) rounds 9-15, (d) rounds 16-22.

—	Single occurrence of CC	━	Four occurrences of CC
═	Two occurrences of CC	▬	Five occurrences of CC
—	Three occurrences of CC	■	Six or more occurrences of CC

We turn next to the time courses of the performances. The time courses of overall C frequencies under the two conditions are shown in Figure 4. The C frequencies appear to rise slightly under the FEEDBACK condition but actually decline slightly under the NO FEEDBACK condition.

Time courses of CC and DD frequencies under the two conditions are shown in Figures 5 and 6. Here, we see that FEEDBACK appears to have the largest effect on increasing reciprocal cooperation (lock-ins on CC) but hardly any on increasing reciprocal noncooperation (lock-ins on DD).

Figures 7 and 8 are pictorial representations of the evolution of reciprocated cooperation in two seven-person groups representative of the FEEDBACK and the NO FEEDBACK conditions, respectively.

Initially, that is, during the first round, four of the twenty-one pairs of the seven-person group cooperated: (1, 2), (2, 3), (2, 5), and (3, 5). (Cf. Figure

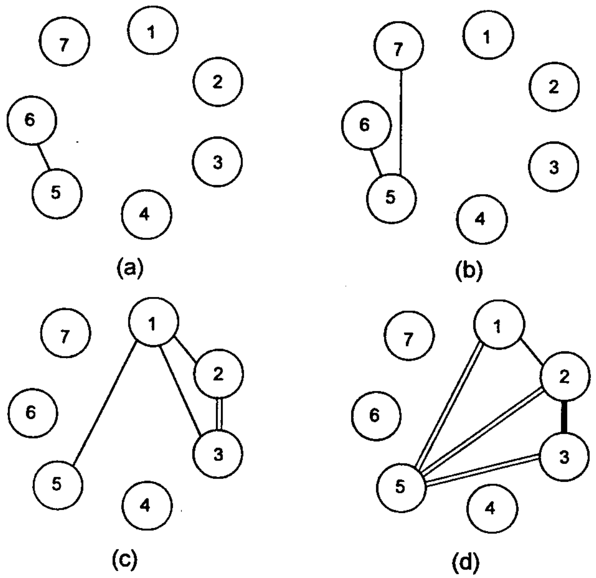

Figure 8: Cooperative links in a group playing under the NO FEEDBACK condition in successive stages: (a) round 1, (b) rounds 2-8, (c) rounds 9-15, (d) rounds 16-22.

─	Single occurrence of CC	▬	Four occurrences of CC
═	Two occurrences of CC	▬	Five occurrences of CC
≡	Three occurrences of CC	▬	Six or more occurrences of CC

7.) In rounds 2-8, the same pairs continued to cooperate repeatedly. In rounds 9-16, player 5 was dropped but (1, 6) produced a cooperative outcome. Finally, player 5 started to cooperate again with players 2 and 7. In the last seven rounds, some degree of cooperation involved every player except 4.

Turning to a selected performance under the NO FEEDBACK condition, we see only slight evidence of evolving cooperation. No lasting lock-ins on CC (represented in Figure 7 by the thickest lines connecting players) occurred. The (partial) evolution of cooperation in the absence of feedback on the past choices of the current player can be explained by the fact that the identity of the current co-player is announced. If this co-player is remembered having cooperated on a previous encounter (with self), the likelihood of cooperating with him or her is enhanced, and this enhancement can produce a (comparatively weak) positive feedback effect.

NOTES

1. In Bern, the instructions were in German.

2. Note that here "round" refers to a sequence of plays in which each subject has played once. In our discussion, "round" means a sequence of plays in which each subject has played once against every other subject in the group, that is, a round robin. Each group played 22 of these round robins.

REFERENCES

Axelrod, R. 1984. *The evolution of cooperation*. New York: Basic Books.

Coleman, J. S. 1988. Social capital in the creation of human capital. *American Journal of Sociology* 94:95-120.

Granovetter, M. 1985. Economic action and social structure: The problem of embeddedness. *American Journal of Sociology* 91:481-510.

Raub, W., and J. Weesie. 1990. Reputation and efficiency in social interactions. An example of feedback effects. *American Journal of Sociology* 96:626-54.

Selten, R. 1978. The chain store paradox. *Theory and Decision* 9:127-59.

The extent to which trust prevails can be measured by the subjective probability with which an agent expects another one to act in desired ways. An agent's trust in other agents forms during repeated social interactions that typically have the structure of an elementary game of trust. The process of trust formation in such interactions can be described by a reputation function. It is argued that in view of real-world processes of trust formation, any adequate reputation function must satisfy certain conditions. A simple model conforming to these conditions is presented. Analyzing this example, it is shown that there is a cooperative Nash equilibrium in a trust supergame, which is in accordance with the basic conditions of realistic trust formation. However, it is also proved that no process of trust formation, can be reasonably similar to real-world mechanisms and at the same time lead to subgame perfect equilibria in a trust supergame.

Trust and Strategic Rationality

BERND LAHNO
University of Duisburg, Germany

As game-theoretic models of problematic social situations show,[1] individual and so-called collective rationality may clash. The Prisoner's Dilemma is, of course, the paradigm example. If the prisoners could rationally behave trustworthily and rationally trust each other, the Pareto superior result could be reached. However, rationally they can do neither.

As has been argued by several authors (e.g., Dasgupta 1988; Kreps 1990), a simple game of perfect information may be used to illustrate the essential aspects and fundamental problems of cooperation based on trust. The general form of this game is given by the game tree of Figure 1.

I shall henceforth refer to this paradigmatic model of a very general and fundamental class of decision situations as the *elementary trust game*. This game may be interpreted as modeling the basic features of any successive

Author's Note: *An earlier German version of this article was written during a sponsored visit at the ISCORE, Department of Sociology, Utrecht University. I would like to thank Hartmut Kliemt for many valuable comments and for helping me with the English translation. I am also indebted to two anonymous referees.*

RATIONALITY AND SOCIETY, Vol. 7 No. 4, October 1995 442-464