

Social exchange and reciprocity: confusion or a heuristic?

Toko Kiyonari, Shigehito Tanida, Toshio Yamagishi*

Graduate School of Letters, Hokkaido University, N10 W7 Kita-ku, Sapporo 060-0810, Japan

Received 26 April 2000; accepted 23 July 2000

Abstract

We propose that a “social exchange heuristic” is as important as the cheater detection mechanism for attaining mutual cooperation in social exchange. The social exchange heuristic prompts people to perceive a mixed-motive situation, such as the Prisoner’s Dilemma (PD), as an Assurance Game (AG) situation in which cooperation is a personally better choice than defection insofar as the partner is cooperating as well. We demonstrate the operation of the social exchange heuristic through a comparison of the ordinary one-shot, simultaneous PD with the one-shot, sequential PD. Participants in the current experiments, involving a total of 261 volunteers, committed a logical error in the direction of favoring mutual cooperation as the situation involved more serious consequences. This result strongly suggests the operation of a domain specific “bias” that encourages pursuit of mutual cooperation in social exchange. © 2000 Elsevier Science Inc. All rights reserved.

Keywords: Prisoner’s dilemma; Social exchange; Heuristic; Reciprocity

1. Social exchange and cheater detection

According to Cosmides (1989) and Cosmides and Tooby (1989, 1992), the history of human evolution has produced domain-specific cognitive modules that are suited for solving problems that humans encounter in social exchange. A problem that Cosmides and Tooby consider central for the successful adaptation of humans is how to promote mutual cooperation in social exchanges, of which the typical (logical) example is Prisoner’s Dilemma (PD) or similar, mixed-motive interdependent situations. How to promote mutual cooperation in PD or Social Dilemmas (SD or n -person version of PD) has been the target of many research endeavors in

* Corresponding author.

E-mail address: toshio@let.hokudai.ac.jp (T. Yamagishi).

various disciplines in the social and biological sciences. Biologists have proposed two answers to the question of how cooperation is possible among self-interested organisms. One answer is *kin selection* (Hamilton, 1964), and the other is *reciprocal altruism* (Trivers, 1971).

Social scientists are naturally interested in the second answer since they are mostly concerned with cooperation among non-relatives. We are also interested in the promotion of mutual cooperation among non-relatives, and thus, our focus in this paper is on the second answer. The idea that reciprocal altruism can explain cooperation in PD games has been proposed by game theorists and social scientists under the rubric of tit-for-tat (TFT) strategy (e.g., Axelrod, 1984). A series of computer simulations conducted by Axelrod (1984) demonstrated that the TFT strategy outperformed all the other strategies in a computer-simulated tournament of strategies. In additional computer simulations that Axelrod conducted, TFT was demonstrated to evolve — i.e., to replace poorly performing strategies and spread to most members of the group. His computer simulations further demonstrated that the TFT strategy is evolutionarily stable against a broad range of alternatives.

Based on these earlier findings and theorizing, Cosmides and Tooby speculated that humans have acquired, through the history of evolution, domain-specific cognitive modules that prompt people to engage in reciprocal altruism — or, to apply a TFT strategy — when they face social exchange situations. Engaging in reciprocal altruism, or applying the TFT strategy in social exchange requires, at the minimum, an ability to distinguish cooperators from defectors. Cosmides and Tooby thus reasoned that humans should have acquired through evolution a cognitive module to be activated in social exchange situations that, once activated, directs our attention to detecting defectors. Given the importance of promotion of mutual cooperation for humans, and given the need to distinguish defectors from cooperators for successfully applying the TFT strategy, it is highly plausible, they reasoned, that humans have evolved a cognitive module specific to the social exchange domain for processing information relevant to distinguishing cooperators from defectors.

We accept Cosmides and Tooby's argument that social exchange has been of central importance in human evolutionary history, and that the achievement of mutual cooperation in social exchange is one of the most important adaptive tasks humans have faced. However, while Cosmides and Tooby focused their attention on cheater detection, in this paper, we focus our research on another condition for the promotion of mutual cooperation in social exchange.

Our analysis starts with the obvious fact that having cheater detection ability is of no use for the attainment of mutual cooperation unless one is willing to cooperate with other cooperators.¹ Cheater detection ability is useful only for those who do not want to cheat others — that is, for those who have given up on their desire to unilaterally cheat or exploit cooperative partners — in social exchange. This insight that achievement of mutual cooperation in PD requires both willingness to cooperate (or willingness not to unilaterally exploit cooperative partners) and cognitive ability to distinguish defectors from cooperators

¹ This argument is based on the assumption that others also have some cheater detection ability. Given this assumption, cheaters will not be accepted as exchange partners even when they want to find their prays. Cheater detection ability is thus useful only for those who can be accepted by others as exchange partners, i.e., for cooperators.

comes from Pruitt and Kimmel's (1977) goal/expectation theory of cooperation. They reviewed experimental gaming studies, which had accumulated to exceed 1000 by the mid-1970s, and came to the conclusion that achievement of cooperation in PD requires both of the following two elements: (1) transformation of the goal from one-sided pursuit of one's own benefit to attainment of mutual cooperation, and (2) expectation that the partner is also willing to forgo narrowly egoistic pursuit of benefit and, instead, is aiming for mutual cooperation. Cosmides and Tooby's "logic of social exchange" focuses on the "expectation" component while taking for granted the "goal" component (i.e., willingness to cooperate) in Pruitt and Kimmel's goal/expectation theory.

It is obvious that those who desire to exploit cooperative partners cannot achieve mutual cooperation even when they have the ability to distinguish defectors from cooperators. Conversely, those who one-sidedly cooperate without paying any attention to the possibility of being exploited by defecting partners cannot achieve mutual cooperation, either. It is a simple logical truism that in order to successfully promote mutual cooperation in social exchange, we need both the cognitive mechanisms to facilitate transformation of the goal and the mechanisms to distinguish defection from cooperation. It then follows that if we have evolved a cognitive module specialized for cheater detection in social exchange, we should also have evolved another cognitive module specialized for persuading us to abandon the one-sided pursuit of exploiting a cooperative partner. We focus in this paper on the second, and call it the *social exchange heuristic*.

1.1. The social exchange heuristic as the subjective transformation of PD into an Assurance Game (AG)

How does the social exchange heuristic prompt us to abandon desires to exploit partners in social exchange and seek mutual cooperation with them? Our answer to this central question is the subjective transformation of a PD into an AG. That is, humans have a cognitive bias in the information processing of social exchange, according to which they perceive PD-like situations as an AG. In PD, defection is a dominant choice. That is, defection produces an individually better outcome no matter what the choice of the partner is. In contrast, there is no dominant choice in AG. Defection produces an individually better outcome when the partner also defects. However, cooperation produces an individually better outcome when the partner cooperates. With this subjective transformation, people intuitively perceive most mixed-motive incentive structures as ones in which mutual cooperation is personally more desirable — that is, produces personally better outcomes — than defection insofar as the partner also cooperates. Examples of a PD and an AG are presented in Table 1.

It is natural that players of iterated (i.e., repeated between the same partners) PD see the situation as AG. Each constituent game in an iterated format is a PD, but the whole series constitutes an AG. Application of the TFT strategy or reciprocal altruism transforms the game into an AG supergame. (Supergame is the term used in game theory to describe the whole series of iterated games.) However, one-shot PD is a different story. In a one-shot PD, each player makes a decision once, and simultaneously, so there is no possibility that one player's choice affects the partner's choice. TFT strategy is thus

Table 1
Examples of a PD and an AG

An example of a Prisoner's Dilemma			An example of an Assurance Game		
Player B's Choice	Player A's Choice		Player B's Choice	Player A's Choice	
	Cooperation	Defection		Cooperation	Defection
C	2	0	C	2	0
D	3	1	D	1	1

Within each quadrant, representing a joint decision of the two players, player A's payoff is indicated above and to the right of the diagonal, and player B's payoff is represented below and to the left.

useless in a one-shot PD. And yet, experimental evidence shows that players of one-shot PD games often play them as if they were AG; they are subjectively transforming PD into AG.

Direct evidence that players of a one-shot PD game do in fact play it as an AG comes from players themselves who indicate that they prefer the outcome of mutual cooperation to the outcome of unilateral defection. According to Kollock (1997), participants in a vignette-type experiment indicated that they prefer the outcome of mutual cooperation to that of unilateral defection unless the partner is a member of an antagonistic group. Similar evidence is also reported by Watabe, Terai, Hayashi, and Yamagishi (1996). Participants in their experiment played a one-shot PD and then rated desirability of the four outcomes on a seven-point scale. On average, they judged that cooperation would yield a personally more satisfactory outcome than defection when the partner cooperated (6.22 vs. 4.62). On the other hand, they judged that not cooperating would yield a more satisfactory outcome than cooperating when the partner failed to cooperate (3.16 vs. 1.82). The pattern was replicated with American participants by Hayashi, Ostrom, Walker, and Yamagishi (1999); the corresponding means among American participants were (6.01 vs. 5.05) and (3.66 vs. 1.76).²

Table 2 shows the average "satisfaction" scores for the four outcomes. The four outcomes are represented in the table by the combinations of two characters, C and D, the first character indicating the choice of the participant and the second, the choice of the partner. The table also shows the proportion of the participants whose stated preferences match the incentive structure of the AG and that of the PD game. Based on the questionnaire responses, 41% of the participants in Watabe et al.'s (1996) experiment actually played the PD game as an AG (i.e., preferred CC to DC and DD to CD), and only 18% of them played it as a PD game per se (i.e., preferred DC to CC and DD to CD).

² These figures do not provide clues regarding whether the transformation involves extra appeal of the mutual cooperation outcome, aversion toward the unilateral exploitation outcome, or both.

Table 2

Average “satisfaction” scores and preference rankings of the four outcomes in the PD game, and the proportion of participants who played PD as such and as an AG in six experiments conducted by Yamagishi and his colleagues

Source	n	Mean desirability				Mean preference ranking				Proportions			
		CC	DC	CD	DD	CC	DC	CD	DD	Desirability		Ranking	
										PD ^a	AG ^b	PD ^a	AG ^b
Watabe et al. (1996)	148	6.22	4.62	1.82	3.16					0.18	0.41		
Hayashi et al. (1999)	167	6.01	5.05	1.76	3.66					0.28	0.30		
Terai (1995)	81	6.33	4.71	1.94	3.79	1.38	2.46	3.47	2.70	0.22	0.46	0.04	0.40
Kiyonari et al. (1998)	79	6.44	4.82	2.13	3.59	1.43	2.43	3.43	2.68	0.22	0.42	0.30	0.47
Yamagishi and Kosugi (1999)	40	6.18	4.78	2.50	4.08	1.23	2.58	3.49	2.67	0.18	0.43	0.20	0.58
Yamagishi et al. (1999)	90					1.37	2.32	3.60	2.69			0.27	0.51

CC, DC, CD, DD: The first character represents the participant’s own choice and the second character the partner’s choice.

^a Proportion of PD players who preferred CC to DC and DD to CD.

^b Proportion of AG players who preferred DC to CC and DD to CD.

Among the American participants in Hayashi et al.’s (1999) experiment, 30% played the game as an AG and 28% played it as a PD game per se.³

In three other experiments on one-shot PD games conducted by Kiyonari, Yamagishi, and Nakajima (1998); Terai (1995); Yamagishi, Kikuchi, and Kosugi (1999), and Yamagishi and Kosugi (1999), participants expressed their preferences for the four outcomes by assigning preference orders to the four outcomes in addition to evaluating “satisfaction” with each outcome separately. The averages of those preference rankings were consistent with the average “satisfaction” scores of the four outcomes, indicating that on average participants in these experiments played the game as an AG. The proportion of the participants who played the game as AG (hereafter called “AG players”) ranged from 40% to 58%, and the proportion of the participants who played it as a PD game (hereafter called “PD players”) ranged from 4% to 30% (see Table 2). Finally, participants in the sixth experiment (Yamagishi et al., 1999) shown in Table 2 expressed their preferences for the four outcomes only by ranking the desirability of the four outcomes. The result of this experiment was also consistent with those in the previous experiments, showing that more participants played the game as an AG than as PD.

³ The proportions of the PD players and the AG players do not add up to 100% since some participants have other preference orders. The most frequently observed preference order besides PD or AG was CC=DC and DD>CD (13.5% among Japanese participants and 21.0% among American participants). Focusing only on the comparison between CC and DC, 29 Japanese and 20 American participants who did not qualify either as PD or AG players preferred CC to DC, whereas only two Japanese and six American participants preferred DC to CC. Twenty-seven Japanese and forty-four American participants were indifferent between CC and DC.

The preferences for the four outcomes stated by the participants in the above experiments consistently indicate that a majority of participants played the one-shot PD game as an AG. However, it is possible that the stated preferences are just “lip service” and do not reveal true preferences. Experiments by Hayashi et al. (1999) and Watabe et al. (1996) provide a safeguard against this criticism, however. Those experiments included two “sequential-game” conditions in which participants were first informed of the partner’s choice and then decided whether to cooperate or defect. In one of these conditions, participants were informed that the partner had already made a decision to cooperate. In the other condition, they were informed that the partner had already made a decision to defect. The overwhelming majority of the participants (20 of the 23 Japanese participants in Watabe et al.’s experiment, and all of 13 American participants in Hayashi et al.’s experiment) in the second condition defected, as everyone would expect. But what about the first condition? If the participants’ “true” preferences were consistent with the monetary value of the four outcomes, they would have defected in the first condition as well. Quite contrary to this expectation, the majority of the participants in this condition who knew that the partner had already made a decision to cooperate (75% or 15 of the 20 Japanese participants in Watabe et al.’s experiment, and 61% or 11 of the 18 American participants in Hayashi et al.’s experiment) cooperated. A replication of these experiments with Korean participants by Cho and Choi (1999) produced a similar pattern: When the first player defected, none of 10 cooperated, and when the first player cooperated, 73% or 8 of 11 cooperated. Those cooperation rates were much higher than those in the simultaneous-game condition (56% among Japanese participants, 36% among American participants, and 46% among Korean participants). Furthermore, practically all (9 of the 10 Japanese and all of the 6 American) participants in this condition who expressed the preference pattern of an AG (AG players) cooperated. In contrast, the overwhelming majority of the PD players (two of the three Japanese and all of the three American participants) in this condition defected. These results show (1) that many of the participants play PD game as an AG, and (2) that their actual behavior is consistent with their stated preferences.

1.2. Confusion or a heuristic?

The social exchange heuristic, we claim, is triggered by the construal of the situation as one involving social exchange, and once triggered, subjectively transforms the nature of the exchange, resulting in the perception of the situation as an AG. With the transformed perception of the situation, people come to intuitively believe in the desirability of mutual cooperation. Activation of the social exchange heuristic motivates people to seek mutual cooperation. We conducted an experiment to demonstrate the operation of this social exchange heuristic. In the following experiment, we focused our analysis on the one-shot version of PD, instead of the iterated or repeated version of it. This is because we wanted to demonstrate that the social exchange heuristic operates even in the one-shot situation. In iterated PD in which TFT objectively transforms the whole series of games into an AG supergame, it is not surprising to find that players prefer mutual cooperation and seek to establish mutual cooperation.

Our aim is to demonstrate that the transformation occurs at the subjective level even in one-shot PD.

Faced with the aforementioned experimental findings showing the operation of the social exchange heuristic, proponents of the rational choice approach to the analysis of social exchange often claim that cooperation in one-shot PD, either in the dyadic form or in the n -person form that is called a public goods provision game or a SD game, is a result of confusion on the part of the participants (cf., Andreoni, 1995). According to this “confusion hypothesis” (Andreoni, 1995), what we call the social exchange heuristic is in fact an expression of participants’ confusion: They fail to understand the incentive structure and the one-shot nature of the game that makes defection a dominant choice. Logically, what we call a social exchange heuristic — i.e., subjective transformation of PD into AG — produces an error. We consider this “error” to be an important design feature of human cognitive functioning that makes mutual cooperation possible in social exchange, whereas rational choice theorists regard mutual cooperation in social exchange as an “accidental” result of confusion.

The main purpose of this study is to examine the validity of these two interpretations of cooperation in one-shot PD — confusion vs. heuristic. The study is based on the reasoning that how seriously participants take their choices in PD provides a critical test between the confusion interpretation and the social exchange heuristic interpretation. We argue that the social exchange heuristic, once triggered, transforms PD into AG. We thus predict that cooperation rates will be higher in the situation in which the social exchange heuristic is more likely to be triggered than in the situation in which it is less likely to be triggered. We further argue that the perception of the situation as one involving social exchange, thus triggering the social exchange heuristic, requires a realistic sense of exchange in which participants’ decisions produce significant outcomes for themselves and their partners. Conversely, the experimental situation is not likely to trigger the social exchange heuristic when the participants perceive it as something involving trivial outcomes or as a kind of game in which the only goal is to win. We thus predict that the cooperation rate in one-shot PD will be higher when participants perceive the experimental task as involving serious consequences than when they see no serious consequences of their decisions for themselves or for the partner.

Hypothesis 1: Realistic sense of exchange (i.e., the degree to which players’ decisions involve significant outcomes for themselves and for the partner) will improve players’ cooperation rate.

The confusion hypothesis (Andreoni, 1995), on the other hand, would predict the opposite effect of the significance of outcomes. Players will pay more attention to the nature of incentives and the one-shot nature of the game, and thus would be less confused, when the outcomes are more serious. Players are expected to be less cooperative when they are less confused and thus are more likely to make rational decisions. The following alternative hypothesis is thus derived from the confusion hypothesis.

Alternative Hypothesis 1: Realistic sense of exchange will depress players’ cooperation rate.

1.3. Simultaneous vs. sequential games

The second purpose of this study is to demonstrate the operation of the social exchange heuristic through a comparison of a simultaneous game and a sequential game. We argued earlier that the enhanced cooperation among the second players in a sequential PD was due to the operation of the social exchange heuristic. The sequential nature of the game promotes the sense of contingency between the two players' choices. For the following experiment, we predict that the social exchange heuristic is more likely to be triggered among players of a sequential game than among players of a simultaneous game. Hence, the second hypothesis to be tested in the following experiment.

Hypothesis 2: The cooperation rate among the second players in the sequential game, who know that the first player has decided to cooperate, will be higher than that among the players of the simultaneous game.

We believe that a prediction in the opposite direction would be derived with respect to this comparison from the confusion hypothesis, since the second players face a less complex decision task than do players of a simultaneous game (cf., Shafir & Tversky, 1992). They need to compare only two outcomes instead of four, and they need not to speculate on the choice of the first player. Furthermore, there is no room for "illusion of control," or the belief that one's own behavior would affect the partner's choice even in one-shot PD, since it is clear that they cannot affect the choice of the partner whose decision has already been made. The following alternative hypothesis would thus be derived.

Alternative Hypothesis 2: The cooperation rate among the second players in the sequential game, who know that the first player has cooperated, will be lower than that among the players of the simultaneous game.

Alternative Hypothesis 2 is the one proposed and supported by Shafir and Tversky (1992). They reasoned that the illusion of control would be greatly reduced when the game is played sequentially rather than simultaneously. More specifically, the illusion of control would be greatly reduced for the second player in a sequentially played PD who knows that the first player has already made his or her decision and thus, he or she (the second player) cannot affect the first player's decision. They thus predicted that the cooperation rate among the second players who were informed that the first player had already made the decision to cooperate would be lower than the cooperation rate in the simultaneous game in which two players made their decisions simultaneously. Their experimental results supported this prediction; their participants cooperated at a lower rate when they were informed that the partner had already made the decision to cooperate than when the PD was played simultaneously.

This finding by Shafir and Tversky (1992) contradicts Hypothesis 2 presented above and the experimental findings by Cho and Choi (1999); Hayashi et al. (1999), and Watabe et al. (1996) presented earlier. It also contradicts the findings of Morris, Sim, and Grotto's (1998) experiment in which participants cooperated at a higher level in the sequential game in which they knew that the first player had already cooperated than in the simultaneous game. In a slightly different

context, McCabe, Smith, and LePore (2000) report experimental finding showing that players of a PD game cooperated at a higher level when the game was presented in the extensive form than in the normal (i.e., matrix) form. When the game was played in the extensive form, it was played sequentially; when the game was played in the normal form, it was played simultaneously.

What explains these contradictory findings? The answer we are proposing here is related to the lack of realistic sense of exchange for the participants in the Shafir and Tversky (1992) experiment; we hypothesize that the lack of a realistic sense of exchange in that experiment resulted in a failure to trigger the social exchange heuristic. Shafir and Tversky's participants played the experimental game 40 times, and accumulated points over the 40 outcomes. The point total was converted into real money only at the conclusion of the experiment — far removed from the time of decision — according to a formula that was not specified to participants. Facing such tedious and tiring tasks, and a very weak outcome of each decision (a few cents per decision in reality, although participants did not even know how much they could make or lose as a result of their decisions), it is likely that the participants did not take each decision very seriously. Furthermore, the nature of the random matching with a new partner in each decision round was likely to prevent them from perceiving the situation as a social exchange situation, the goal of which is the attainment of mutual cooperation. In sum, Shafir and Tversky's experimental design with many rounds of experimental games with a randomly selected new partner in each trial is the least encouraging environment for the social exchange heuristic to be triggered. In contrast, Hayashi et al.'s (1999) and Watabe et al.'s (1996) participants played a PD game once, and their reward was directly dependent on the outcome of their decisions in that single PD. Naturally, they took their tasks seriously, at least much more seriously than did the participants in Shafir and Tversky's experiment.

Our interpretation of Shafir and Tversky's (1992) findings presented above leads us to the third hypothesis to be tested in the experiment. We predict that Shafir and Tversky's finding that the cooperation rate is higher among the simultaneous players than the second players with a cooperative first player will obtain only when the social exchange heuristic is not likely to be triggered, that is, only when outcomes are trivial and thus, a realistic sense of exchange is not felt by the player; the other pattern predicted in Hypothesis 2 will emerge when the game outcomes are non-trivial and a realistic sense of exchange is felt by the players.

Hypothesis 3: The enhanced cooperation rate among the second players predicted in Hypothesis 2 will exist only when the game outcomes are non-trivial.

The confusion hypothesis would predict a different pattern. Confusion is expected to be more serious among the simultaneous players than among the second players, especially when the game outcomes are trivial. When the game outcomes are more serious, players would more carefully examine the nature of the game and thus, would be less likely to be confused. Thus, the enhanced cooperation rate among the simultaneous players predicted in Alternative Hypothesis 2 will be more prominent when the game outcomes are trivial than when the game outcomes are non-trivial.

Alternative Hypothesis 3: The enhanced cooperation rate among the simultaneous players predicted in Alternative Hypothesis 2 will exist only when the game outcomes are trivial.

1.4. The first player in the sequential game

In the following set of experiments, we varied the level of the realistic sense of exchange or the degree to which the outcomes of the participants' decisions have relevance to their own and their partner's welfare. Specifically, we compared two types of experiment: a full experiment in which participants' payments were dependent on their decisions and a vignette experiment in which participants' decisions did not have any tangible consequences for them. In the vignette experiment, participants imagined that they had been in an experiment described in a scenario instead of actually participating in an experiment.

In addition to the two conditions — the simultaneous-game condition and the second-player condition — we added a third condition: the first-player condition. The participant in this condition was told that he or she would make the decision first (without knowing the second player's choice); the second player would make a decision with the knowledge of the first player's choice. This is a condition in which the illusion of control is no longer an illusion. As shown in Hayashi et al.'s (1999) and Watabe et al.'s (1996) findings, the second player's choice is greatly affected by the first player's. In their experiment, the second player cooperated at a much higher level when the first player cooperated than when the first player defected. The confusion approach cannot predict if the cooperation rate among the first player will be higher or lower than the simultaneous players or the second players. We predict, based on the social heuristic approach, that the cooperation rate in the first-player condition will go hand in hand with that in the second-player condition. This is because, once the social exchange heuristic is triggered, the first player expects the second player to reciprocate. The situation that triggers the social exchange heuristic will enhance both the second player's reciprocation of the first player's cooperation and the first player's expectation of reciprocation from the second player. With the expectation of reciprocation from the second player, it is a subjectively better choice for the first player to cooperate than to defect. Thus, we derive the following two additional hypotheses.

Hypothesis 2f (for the first player): The cooperation rate among the first players in the sequential game will be higher than that among the players of the simultaneous game.

Hypothesis 3f (for the first player): The enhanced cooperation rate among the first players predicted in Hypothesis 2f will exist only when the game outcomes are non-trivial.

2. Method

In order to manipulate the realistic sense of exchange — i.e., seriousness of participants' decisions for themselves and for the partner — we ran the experiment using two different formats. First, the experiment was run as a fully fledged experiment in which the participants played a PD game once and were paid exactly the amount specified in the

Table 3
The payoff matrix of the PD used in the experiment

Your choice	The other person's choice			
	K		P	
L	You get ¥1200	The other person gets ¥1200	You get ¥0	The other person gets ¥1800
S	You get ¥1800	The other person gets ¥0	You get ¥600	The other person gets ¥600

payoff matrix.⁴ In that format, the three player-type conditions — the simultaneous-player condition, the second-player condition, and the first-player condition — constituted a between-subjects factor. Each participant played a PD game only once in only one of the three conditions. Participants were brought to the laboratory, and were given instructions and plenty of time to make a decision with the full knowledge that their earnings would be determined by the outcome of their decisions. The second experimental format involved the use of vignettes. They were told to imagine that they had been participating in the experiment described in the vignette, and to decide whether they would have cooperated or defected if they had been in the experiment. The more detailed description of each format will follow.

2.1. Full experiment

A total of 149 participants (108 males and 41 females) were recruited from a participant pool of about 1500 at a major Japanese national university. The participant pool had been created by soliciting potential participants from various introductory courses on campus. Registration for the participant pool was completely voluntary. Monetary incentives were emphasized for soliciting potential participants.

Several participants were contacted through telephone calls for a particular experimental session. As they individually arrived at the reception desk set up in the entrance lobby area of the building, they were individually given an ID card and were told that they would be identified by the ID number throughout the experiment to ensure anonymity. They were then individually escorted to the laboratory and to their own compartment in the laboratory. Thus, they did not have a chance to meet other participants unless more than one participant showed up at the reception desk at the same time. The participants then read the instructions in the isolation of their own compartment. The instructions first explained the nature of the PD game used in the experiment. The payoff matrix shown in Table 3 was used in the instructions. They were given a short quiz after the instructions about the nature of the game to ensure their understanding of the incentive structure. The experimenter checked their answers and provided further clarifications when needed. Participants were warned in the instructions that the game may be played in several different ways. The manipulation of the three conditions was carried out with the second

⁴ Their actual payment was not based on the actual outcome of the PD game. According to the payoff matrix shown in Table 3, the sucker's payoff is zero, but we paid 1200 yen to ensure that the participants earned money from the experiment. That is, they were made to believe that the partner had cooperated.

round of instructions that were given to the participants after their answers to the quiz had been checked.

In the simultaneous-game condition ($n=48$), participants were provided with the same payoff matrix again, were told that the partner was also making his or her decision, and were asked to choose between L (cooperation) and S (defection). In the first-player condition ($n=51$), they were told that they would make the decision first. They were further told that the second player would be informed of their decision before making his or her decision. Finally, in the second-player condition ($n=50$), participants were first informed that the other person (the first player) was making the decision and were asked to wait until he or she made the decision. Then, they were told that the other person had chosen K (cooperation), and then were asked to make their own decision.

Participants made the decision by circling either L or S on the decision sheet, and placed their decision sheet in an envelope and handed it to the experimenter. Then, they filled out a post-experimental questionnaire, placed it in an envelope, and again handed it to the experimenter. Another experimenter located in another room calculated payments to the participants based on their decisions, placed them in individual envelopes, and gave it to the first experimenter. The first experimenter gave the envelope with the payment to the participant. Thus, the first experimenter who met the participant did not know the participant's decision or how much he or she made. The second experimenter knew who made which decision, but identified the participants only by their ID number. Thus, anonymity of the participant's decision was completely maintained. How anonymity would be maintained was explained to the participants in advance. The whole experiment, including a rather lengthy post-experimental questionnaire, some items of which were used for another study, took between 50 and 80 min.

2.2. *Vignette experiment*

A total of 112 students participated in the experiment with vignettes. The total of 112 participants consisted of two groups. One group of participants ($n=56$; 39 males and 17 females) was recruited from the same participant pool as in the full experiment described above. They were brought to the laboratory in the same way as in the full experiment. They were paid a flat fee of 1000 yen for their participation in the experiment. The second group of participants consisted of 56 students at another national university in Japan taking an introductory psychology class. (Sexes of the second group of participants were not recorded.) They were asked by the instructor at the beginning of a class session to donate 15 min of their class time for a psychology experiment. The first group of paid participants were provided with the instructions for the full experiment and were then asked to indicate if they would have chosen L (cooperation) or S (defection) if they had been in that experiment. The instructions were distributed and responses were collected individually, and the participants read the instructions and gave responses in their individual compartments. The second group of participants received the same set of instructions. For the second group, however, instructions were distributed in the classroom and the instructor collected the responses after 15 min.

In the instructions, participants were told that they would read the instructions that were actually used in one experiment, and were asked to imagine that they had been

participating in the actually conducted experiment and to indicate (with the same response format actually used in the full experiment) which of the two alternatives, L or S, they would have chosen if they had participated in the experiment. Each participant of either group was provided with the instructions for all three conditions — simultaneous-game condition, the first-player condition, or the second-player condition — and was asked to indicate his or her imaginary decisions in all three conditions. The participants who were assigned to the second-player condition were told that the partner had already decided to choose K (cooperation).

There were two between-subject conditions in the vignette experiment, the “money” condition and the “score” condition. Fifty-eight participants in the money condition were shown the payoff matrix expressed in terms of monetary value (i.e., yen) as in the original, full experiment. The other 54 participants were shown the same payoff matrix, but there was no mention of money. They were simply told that the numbers shown in the table represented scores and the participants of the original experiment had been instructed to maximize their own scores. Thus, the three game conditions constituted a within-subjects factor, and the money vs. score conditions constituted a between-subjects factor.

Between the two conditions in the vignette experiment — the money condition and the score condition — the decision task in the score condition would be more trivial than that in the money condition. Even when the participants took the instruction seriously and imagined that they were in fact participating in the experiment, no serious consequences were expected in the score condition where the outcomes were just abstract scores. Thus, a realistic sense of exchange is considered to be felt most strongly in the full experiment, followed by the money condition of the vignette experiment, and least strongly in the score condition of the vignette experiment.

3. Findings

3.1. Hypothesis 1 vs. Alternative Hypothesis 1

As shown in Table 4, which reports cooperation rates in the three game conditions in the full experiment and the vignette experiment, realistic sense of exchange was positively related to the level of cooperation.⁵ In all of the three game conditions, the cooperation rate was highest in the full experiment, followed by the vignette experiment with money, and then by the vignette experiment with score. Since the design differs between the full experiment and the vignette experiment — the former involving a between-subject factor and the latter a within-subject factor for the manipulation of the player type — we decided to use only the first response in the vignette experiment for the following statistical analyses so that we could analyze the data with a fully crossed, between-subject design (experiment type \times player

⁵ In the analysis presented below, we combined two groups of participants in the vignette experiment, those who were paid a flat fee and those who were not paid at all. A preliminary analysis of the first choice used in the following statistical analysis revealed no significant differences between the two groups.

Table 4
Cooperation rates in the full experiment and the vignette experiment

	Payoffs expressed in	Game condition		
		Simultaneous player	Second player	First player
Full experiment	Money ($n = 149$)	37.5% ($n = 48$)	62.0% ($n = 50$)	58.8% ($n = 51$)
Vignette experiment (all responses)	Total ($n = 112$)	26.8%	35.5% ($n = 110$) ^a	31.3%
	Money ($n = 58$)	29.3%	53.6% ($n = 56$) ^a	41.4%
	Score ($n = 54$)	24.1%	16.7%	20.4%
Vignette experiment (first response)	Total ($n = 112$)	26.5% ($n = 34$)	32.4% ($n = 34$) ^a	35.7% ($n = 42$)
	Money ($n = 56$)	23.5% ($n = 17$)	52.6% ($n = 19$) ^a	55.0% ($n = 20$)
	Score ($n = 54$)	29.4% ($n = 17$)	6.7% ($n = 15$)	18.2% ($n = 22$)

^a There were two missing responses in the second player condition (in the second group of participants).

type).⁶ The cooperation rates for the first responses are also reported in Table 4 in the row called “vignette experiment (first response).” The first responses were mostly consistent with those including all three within-subject responses reported in the row for “vignette experiment (all responses).” The main effect of the experiment type (full experiment, vignette with money, and vignette with score) was highly significant in this analysis, $\chi^2(2) = 15.23$, $P < .001$. This result clearly supports Hypothesis 1 and refutes Alternative Hypothesis 1. Neither the main effect of player type nor the interaction effect was significant.

3.2. Hypotheses 2 and 3 vs. Alternative Hypotheses 2 and 3

The cooperation rate among the second players (who faced a cooperative first player) was higher in the full experiment than that among the simultaneous players (62.0% vs. 37.5%), and the difference was significant, $\chi^2(1) = 5.88$, $P < .05$. The difference was in the same direction in the vignette experiment with money (53.6% vs. 29.3% with all responses; 52.6% vs. 23.5% with only the first responses), $\chi^2(1) = 3.20$, $P < .05$. In the vignette experiment with score, however, the simultaneous players cooperated more than the second players (24.1% among the simultaneous players vs. 16.7% among the second players, with all responses; 29.4% vs. 6.7% with only the first responses), though the difference was not significant, $\chi^2(1) = 2.71$, ns. The experiment type by game type interaction effect was significant, $\chi^2(2) = 4.98$, $P < .05$. Interestingly, Shafir and Tversky’s (1992) prediction that the simultaneous players would cooperate at a higher level than the second players did materialize, but only in the vignette experiment with score, where the outcome of the PD game was truly trivial; participants acted rationally only when they were faced with a trivial decision task. These results clearly support Hypotheses 2 and 3 and refute Alternative Hypotheses 2 and 3.

⁶ The order of the within-subject factor was randomized in the vignette experiment, and thus we can treat the first response as a between-subject factor.

3.3. Hypotheses 2f and 3f

As predicted, the contrast between the first players and the simultaneous players was comparable with that between the second players and the simultaneous players. The first players were more cooperative than the simultaneous players in the full experiment (58.8% vs. 37.5%), $\chi^2(1)=4.50$, $P<.05$, and in the vignette experiment with money (41.4% vs. 29.3% with all responses; 55.0% vs. 23.5% with only the first responses), $\chi^2(1)=3.78$, $P<.05$. In the vignette experiment with score, the pattern was reversed and the first players were less cooperative than the simultaneous players (20.4% vs. 24.1% with all responses; 18.2% vs. 29.4% with only the first responses), although the difference was not significant, $\chi^2(1)=0.68$, ns. The experiment type \times game type interaction effect was not significant, however, $\chi^2(2)=4.04$, ns.

4. Discussion

The overall message of the experimental results is clear. The majority of participants behaved in a non-rational, reciprocal manner when they faced the PD situation with serious consequences, whereas the only time they behaved “rationally” was when they faced a truly trivial decision task (i.e., in the vignette experiment with a payoff matrix expressed in scores, not in yen). Confusion cannot explain cooperation of the second player who knows that the first player has already cooperated. Furthermore, it cannot explain why the second player cooperated more when faced with a serious decision task than when faced with a trivial decision task. It should thus be concluded that the “illusion of control” — the sense that one’s own behavior affects the partner’s behavior — is something that has a root not in simple confusion but in a more substantial human cognitive mechanism. We postulate the mechanism to be the social exchange heuristic. What we call the social exchange heuristic can be conceptualized in several ways. What we have adopted here is a rather cognitive approach to conceptualizing it; we conceptualized it as a subjective transformation of the incentive structure. Another way of conceptualizing it would be more motivational. For example, the participants’ behavior in the current experiments can be interpreted in terms of inequity aversion. We suspect that these are the two sides of a same coin rather than alternative explanations, though careful research is needed to clarify the relationship between the two approaches.

It is true that expecting one’s own behavior to affect the partner’s behavior in one-shot PD is a logical error. The issue is why so many participants make such an error, and why they make such an error more often when they face a serious PD with substantial monetary outcomes at stake than when they face a trivial PD with nothing substantial at stake. The idea of the social exchange heuristic is based on the assumption that making such an error is more adaptive in social exchange situations than making the logically correct decisions. It is true that the one who makes such an error in a PD experiment loses an opportunity to make additional money. However, the consequence of missing the opportunity to attain mutual cooperation in a real life social exchange would often be much more serious than missing a few extra dollars in a one-shot PD experiment. The social exchange heuristic

“biases” us to seek mutual cooperation in social exchange, diverting us from attempts to exploit exchange partners. Whether this “bias” improves the level of our adaptive success or reduces it depends on the relative importance of attainment of mutual cooperation vis-à-vis one-sided exploitation in social exchange. In the domain of social exchange, logically correct reasoning may not produce as good a consequence as making an error in the direction of helping us attain mutual cooperation. Empirically minded economists and game theorists should pay serious attention to the distinction between simple confusion and heuristics that have adaptive functions.

The findings of our experiment also have implications for evolutionary psychology. If we humans have cognitive modules that help us achieve mutual cooperation in social exchange, they should include both the cheater detection module and what we called the social exchange heuristic. The same message has been voiced in the earlier works of Cosmides and Tooby (1989, 1992), but we feel that the importance of a mechanism to encourage us to seek mutual cooperation has been overshadowed by the spotlight on cheater detection. As discussed in Section 1, having only one of them is totally useless unless we have the other. A research issue that derives from this insight would concern the prediction that people who are more strongly motivated to achieve mutual cooperation (or those whose social exchange heuristic is more powerful) are more apprehensive about distinguishing cooperators from defectors or are in fact better at distinguishing defectors from cooperators. This is a rather counter-intuitive prediction implying that more cooperative and trustful people are more prudent and less gullible than less cooperative and distrustful people are. Kikuchi, Watanabe, and Yamagishi (1997), Yamagishi et al. (1999), and Yamagishi and Kosugi (1999), however, have accumulated experimental evidence demonstrating the validity of this counter-intuitive prediction. They show that high-trusters are better than low-trusters at distinguishing cooperators from defectors in one-shot PD. Those experimental findings suggest that inter-relationships between cognitive mechanisms that incline us to seek mutual cooperation and those used in distinguishing defectors from cooperators in social exchange will provide a fertile ground for future research.

Acknowledgments

The research reported in this paper was supported by scientific research grants from the Japanese Ministry of Education and Culture. We would like to thank Toshikazu Hasegawa, Kai Hiraishi, Alan Miller and Rosemary Hopcroft for their comments on earlier drafts of this paper. We would also like to thank an anonymous reviewer for his/her helpful comments and the editors of this journal for their thoughtful suggestions for improving the paper, especially the writing quality.

References

- Andreoni, J. (1995). Cooperation in public-goods experiments: kindness or confusion? *The American Economic Review*, 85, 891–904.

- Axelrod, R. (1984). *The evolution of cooperation*. New York: Basic Books.
- Cho, K., & Choi, B. (1999). A cross-society study of trust and reciprocity: Korea, Japan and the U.S. A paper presented at the WOW II, Workshop for the Political Theory and Policy Analysis, Indiana University, June 16–19.
- Cosmides, L. (1989). The logic of social exchange: has natural selection shaped how humans reason? Studies with the Wason selection task. *Cognition*, 31, 187–276.
- Cosmides, L., & Tooby, J. (1989). Evolutionary psychology and the generation of culture: Part II. A computational theory of social exchange. *Ethology and Sociobiology*, 10, 51–97.
- Cosmides, L., & Tooby, J. (1992). Cognitive adaptations for social exchange. In: J. H. Barkow, L. Cosmides, & J. Tooby (Eds.), *The adapted mind: evolutionary psychology and the generation of culture* (pp. 163–228). New York: Oxford Univ. Press.
- Hamilton, W. D. (1964). The genetical evolution of social behaviour: Parts I, II. *Journal of Theoretical Biology*, 7, 1–52.
- Hayashi, N., Ostrom, E., Walker, J., & Yamagishi, T. (1999). Reciprocity, trust, and the sense of control: a cross-societal study. *Rationality and Society*, 11, 27–46.
- Kikuchi, M., Watanabe, Y., & Yamagishi, T. (1997). Judgment accuracy of other's trustworthiness and general trust: an experimental study. *Japanese Journal of Experimental Social Psychology*, 37, 23–36 (In Japanese with an English abstract).
- Kiyonari, T., Yamagishi, T., & Nakajima, T. (1998). Assurance of security and production of trust. In *Proceedings of the 46th annual meetings of the Japanese group dynamics association* (pp. 292–293).
- Kollock, P. (1997). Transforming social dilemmas: group identity and cooperation. In: P. Danielson (Ed.), *Modeling rational and moral agents* (pp. 186–210). Oxford: Oxford Univ. Press.
- McCabe, K., Smith, V., & LePore, M. (2000). Intentionality detection and “mindreading”: why does game form matter? *Proceedings of the National Academy of Science of the United States of America*, 97, 4404–4409.
- Morris, M. W., Sim, W. M., & Giroto, V. (1998). Distinguishing sources of cooperation in the one-round prisoner's dilemma: evidence for cooperative decisions based on the illusion of control. *Journal of Experimental Social Psychology*, 34, 464–512.
- Pruitt, D. G., & Kimmel, M. J. (1977). Twenty years of experimental gaming: critique, synthesis, and suggestions for the future. *Annual Review of Psychology*, 28, 363–392.
- Shafir, E., & Tversky, A. (1992). Thinking through uncertainty: nonconsequential reasoning and choice. *Cognitive Psychology*, 24, 449–474.
- Terai, S. (1995). Cooperation in a prisoner's dilemma and the perception of interdependence. Master's Thesis, Faculty of Letters, Hokkaido University, Sapporo, Japan. (In Japanese).
- Trivers, R. L. (1971). The evolution of reciprocal altruism. *Quarterly Review of Biology*, 46, 35–57.
- Watabe, M., Terai, S., Hayashi, N., & Yamagishi, T. (1996). Cooperation in the one-shot prisoner's dilemma based on expectations of reciprocity. *Japanese Journal of Experimental Social Psychology*, 36, 183–196 (In Japanese with an English abstract).
- Yamagishi, T., Kikuchi, M., & Kosugi, M. (1999). Trust, gullibility and social intelligence. *Asian Journal of Social Psychology*, 2, 145–161.
- Yamagishi, T., & Kosugi, M. (1999). Character detection in social exchange. *Cognitive Studies*, 6, 179–190 (In Japanese with an English abstract).