ELSEVIER

Journal of Economic Behavior & Organization
Vol. 45 (2001) 1–17

JOURNAL OF
Economic Behavior
& Organization

www.elsevier.com/locate/econbase

Contingent social utility in the prisoners' dilemma

Robert Gibbons [a,*], Leaf Van Boven [b]

[a] *Sloan School of Management, Massacheusetts Institute of Technology, Room E52-590, 50 Memorial Drive, Cambridge, MA 02142-1347, USA*
[b] *Faculty of Commerce, University of British Columbia, Vancouver, BC, Canada*

Received 21 August 1997; received in revised form 9 May 2000; accepted 2 June 2000

Abstract

We examined a central assumption of recent theories: that social utility is contingent on impressions of other people. We manipulated participants' impression of the other player in a prisoners' dilemma. We then measured participants' own preferences in the PD, their estimates of the other player's preferences in the PD, their prediction of the other player's move, and their own move. We hypothesized that the participants' move would maximize their stated preferences, given their prediction of the other player's move, and that participants' preferences would be contingent on their impression of the other player. Results supported both hypotheses and revealed that participants' preferences were contingent more on their estimate of the other player's preferences than on their prediction of the other player's move.

JEL classification: C72 (non-cooperative games); C91 (laboratory experiment, individual behavior)

Keywords: Prisoners' dilemma; Social utility; Contingent utility; Behavioral game theory

1. Introduction

Behavioral decision theory has made important progress by incorporating heuristics and biases that people use in everyday judgement into formal models of single-person decision making (Dawes, 1988; Camerer, 1995; Bazerman, 1998). Some research on negotiations (Neale and Bazerman, 1991) and behavioral game theory (Camerer, 1990) has incorporated these heuristics and biases into analyses of multi-person problems. But the fact that negotiations and games involve more than one person suggests that these literatures may also

* Corresponding author. Tel.: +1-617-253-0283; fax: +1-617-253-2660.
E-mail address: rgibbons@mit.edu (R. Gibbons).

0167-2681/01/$ – see front matter
PII: S0167-2681(00)00170-0

benefit by borrowing other ideas from psychology, such as those on interpersonal perception (Jones, 1990; Gilbert, 1999). For example, the payoffs to the players in a game are a social allocation and so raise issues of social utility: a player's utility in a game may depend not only on her own payoff but also on the payoffs to other players (Kelley and Thibaut, 1978).

The literatures on public-good games and ultimatum games offer two large bodies of evidence regarding social utility. Research on public-good games suggests that some people behave altruistically, contributing more to the public good than would be rational for a purely self-interested person (Ledyard, 1995). In contrast, research on ultimatum games suggests that some people behave spitefully, rejecting small but positive offers that a rational, self-interested person would accept (Camerer and Thaler, 1995). One interpretation of these findings is that some people are always altruistic, others are always spiteful, and still others are always self-interested.

An alternative interpretation of this evidence from public-good and ultimatum games is that people's social utility is contingent rather than fixed. For example, player 1 might care positively about player 2's payoff if 1 thinks 2 is a nice person; but player 1 might care negatively about player 2's payoff if 1 thinks 2 is a jerk. Recent theoretical work by Rabin (1993), Levine (1998), and Sally (1999) incorporates such contingent social utility as a central assumption, although these theories differ in specifying the source of the contingency.

In this paper, we offer an experimental test of whether social utility may be contingent. We present evidence that a player's stated preferences in a prisoners' dilemma may depend on that player's estimate of the other player's preferences. This evidence supports the key assumption of the Rabin, Levine, and Sally theories that social utility is contingent. In the remainder of this introduction, we first provide a brief motivation for and description of these recent theories of contingent social utility. We then describe prior empirical research suggesting that social utility may be contingent. Finally, we outline our study and hypotheses.

1.1. Theories of contingent social utility

In a typical public-good game, each player can make a contribution to a common pool, which is then multiplied by some factor (greater than one but smaller than the number of players) and distributed equally among the players. In such a game, a rational, self-interested player should contribute nothing. But many people do contribute. Summarizing this enormous literature, Ledyard (1995, p. 172) concludes that "hard-nosed game theory cannot explain the data". Some people seem altruistic, at least some of the time.

The evidence from ultimatum games has the opposite flavor. In an ultimatum game, player 1 proposes a division of a fixed and known pie, which player 2 then either accepts or rejects. A rational, self-interested player 2 should accept any positive amount, but many people reject positive offers (Camerer and Thaler, 1995). These results hold even when the game is played in Indonesian villages for stakes equal to three times a month's wages (Cameron, 1999; see also Slonim and Roth, 1998, for similar evidence from the Slovak Republic). Thus, Ledyard's summary that "hard-nosed game theory cannot explain the data" applies

to ultimatum games as well, but for the opposite reason: some people seem spiteful, at least some of the time. [1]

How can these two results be reconciled? It could be that participants are drawn from a distribution of utility functions that range from altruistic through purely self-interested to spiteful. Those who are altruistic contribute in public-good games, whereas those who are self-interested or spiteful do not; and those who are altruistic or self-interested accept small offers in ultimatum games, whereas those who are spiteful do not. But in many public-good games, the fraction of participants who behave altruistically is above 50%, and in some ultimatum games, the fraction who behave spitefully is also above 50%. So this distributional view of social utility may encounter simple accounting problems.

Furthermore, the distributional view of social utility has implications beyond contributions to public goods and rejections of ultimatum offers. For example, the player 1s in ultimatum games must be drawn from the same population as the player 2s, so there should be as many spiteful player 1s as there are spiteful player 2s, and this has implications for the range of offers that should be observed in ultimatum games. Levine (1998) finds that these implications for ultimatum games are not fully consistent with the available data. In short, although the distributional view of social utility may explain some evidence, it appears unable to explain all the relevant evidence on its own.

An alternative view is that players' utility functions are contingent rather than fixed. That is, the same player may be altruistic or self-interested or spiteful, depending on the circumstances. For example, Rabin (1993) models player 1's preference for a particular outcome in a 2×2 game as depending partly on a comparison of that outcome to how player 1 would have fared if player 2 had acted differently. If player 2's alternative action would have reduced player 1's monetary payoff then 1 is grateful and so is willing to forego a modest increase in monetary payoff in order to increase 2's monetary payoff. On the other hand, if player 2's alternative action would have increased player 1's monetary payoff then 1 is angry and so is willing to accept a modest decrease in monetary payoff in order to reduce 2's monetary payoff. In Rabin's model, then, people's social utility is contingent on their prediction of the other player's action, or move. We refer to this as the *moves hypothesis*.

Levine (1998) develops a slightly different model in which player 1's utility function depends on 1's belief about player 2's utility function (rather than on player 1's belief about player 2's action). For example, player 1 places a positive weight on player 2's payoff if 1 believes that 2 places a positive weight on 1's payoffs. Conversely, player 1 places a negative weight on player 2's payoff if 1 believes that 2 places a negative weight on 1's payoffs. In Levine's (1998) model, then, people's social utility is contingent on their estimate of the other player's preferences, or motives. We refer to this as the *motives hypothesis*. [2]

[1] Bolton (1991) summarizes and extends another striking source of evidence on spite: disadvantageous counter-offers in two-stage, alternating-offer bargaining games such as where player 1 proposes to divide an initial pie of US$ 10 into US$ 7 for 1 and US$ 3 for player 2, which player 2 rejects, only to propose a division of the subsequent pie of US$ 5 into US$ 2.50 for each player.

[2] Rabin (1998, p. 22) continues in this vein, noting that "people determine their dispositions toward others according to motives attributed to these others, not solely according to actions taken". In fact, although the specific model developed in Rabin (1993) embodies the moves hypothesis, the general approach taken in that paper (in which player 1's beliefs about player 2's beliefs are arguments in player 1's utility function) can readily be applied to the motives hypothesis.

Finally, Sally (1999) broadens Rabin's and Levine's models by arguing that proximity, familiarity, affection, communication, attractiveness, and other factors can all reduce "social distance", and that reducing social distance increases the extent to which each player cares (positively) about the other's payoff. Sally then applies this model of social utility to the prisoners' dilemma and argues that it is consistent with a great deal of evidence from several literatures.

The main purpose of our investigation is to test the central assumption of these theories that social utility may be contingent. We also offer an initial examination of whether the motives or moves hypothesis offers a more accurate description of the source of contingency in social utility.

1.2. Empirical support for contingent social utility

Theories of contingent social utility are not grounded solely in intuition. Rather, they are based on evidence suggesting that social utility may be contingent. For example, early work in social psychology found that people's decision to cooperate or defect in a prisoners' dilemma was based partly on their prediction of whether the other player would cooperate or defect (Dawes et al., 1977; Kelley and Stahelski, 1970). The positive correlation between people's own action and their prediction of the other player's action may reflect contingent social utility.

More recently, Blount (1995) and Charness (1996) compared player 2's decision in an ultimatum game in response to offers made either by a player 1 or by a disinterested third party or by a random-number generator. Both studies found that player 2 is more likely to accept a small offer made by a disinterested third party or by a random-number generator than the same offer made by a player 1 who stands to gain from the offer. In a similar vein, Pillutla and Murnighan (1996) found that player 2s in an ultimatum game were more likely to accept an offer when they did not know the size of the pie to be divided than when they did know the size of the pie. These studies suggest that people's willingness to reject relatively small offers is contingent on their interpretation of the other player's actions.

Finally, Fehr and colleagues have found evidence of reciprocity in experimental economic environments (Fehr et al., 1993, 1997; Fehr and Falk, 1999). For example, when "firms" choose to pay higher wages, some "workers" choose higher effort (even though the wage has already been paid and there is no future in the relationship). This finding is like the rejections of offers in the ultimatum game, but this is "positive" reciprocity, whereas spiteful rejections in ultimatum games are "negative" reciprocity.

All of these findings suggest that social utility may be contingent, but none of these studies actually measures social utility. In contrast, we directly measure people's social utility and their actions and so can examine the consistency of the two, rather than assuming that social utility is revealed by people's actions (see also Wyer, 1969). Furthermore, we directly manipulate and measure people's estimate of the other player's preferences, and so conduct a direct test of whether social utility may be contingent on one player's impression of the other.

Our closest predecessor is a study by Loewenstein et al. (1989). They asked people to indicate their satisfaction with various outcomes of several hypothetical disputes. The

disputes varied in their context (some were business related and others were not) and in the nature of the relationship (some were positive and others were negative). They found that people generally dislike payoff disparities (disliking unfavorable disparities more than favorable disparities), but that they dislike favorable disparities only in non-business related contexts and only in non-negative relationships. Their findings, indicate that social utility is contingent — in this case, on the nature of the relationship (see also Messick and Sentis, 1985). We study whether social utility is contingent on people's perceptions of the other player in an actual game, rather than in a hypothetical situation, and we study whether people's actions in this game are consistent with their stated preferences.

1.3. Our study

To investigate whether social utility might be contingent, we manipulated whether players held a positive or negative impression of the other player in a prisoners' dilemma. As part of an exercise that we described to participants as a "person perception task," participants were led to believe that their opponents had completed a personality questionnaire in an extremely positive or extremely negative fashion. This extreme manipulation was well-suited to our main goal of studying whether social utility might be contingent, but also raises questions about whether such extreme personalities exist in the real world and about the use of deception in research generally. We address these issues in Section 4.

After this manipulation, we measured four things: participants' stated liking for each possible outcome in the prisoners' dilemma, their estimate of the other player's liking for each possible outcome, their prediction of the other player's move, and their own move. Using these four measures, we tested two hypotheses. First, we predicted that participants' action would maximize their stated preferences given their prediction of the other player's move (the *rational-choice hypothesis*). Second, we predicted that participants' preferences would be contingent on their perception of the other player (the *contingent-utility hypothesis*). We also examined whether participants' preferences would be contingent on their estimate of the other player's preferences more than on their prediction of the other player's move (the *motives-versus-moves hypothesis*).

2. Method

2.1. Participants

Forty-five Cornell University undergraduates from introductory psychology courses participated in our 30 minute study. Participants were told the average earning would be US$ 5 and could be as high as US$ 7. Experimental sessions were conducted in groups of two, four, or six participants. Precautions were taken to ensure that participants within each session were previously unacquainted.

2.2. Procedure

When they arrived at the lab, the experimenter asked participants to complete a "personality questionnaire," explaining that their responses would be used to analyze their behavior

		OTHER PLAYER'S CHOICE	
		Choice A	Choice B
YOUR CHOICE	Choice A	you $5, other player $5	you $1, other player $7
	Choice B	you $7, other player $1	you $2, other player $2

Fig. 1. Prisoners' dilemma game presented to participants.

in the experiment. This bogus questionnaire set the stage for the manipulation of our key independent variable: participants' impression of the other player's personality. The personality questionnaire, which is reproduced in Appendix A, included 20 statements about participants' own personality and general world-view. For example, one statement was, "I believe that the dignity and welfare of others is the most important concern for any society". Another was, "I believe sometimes I must sacrifice the welfare of others for my own benefit". Respondents indicated whether each statement described them by circling *Me*, *Not Me*, or *Neutral*.

Although the personality questionnaire included some items culled from actual psychological inventories, notably the ethics position questionnaire (Forsyth, 1980) and the Machiavellianism questionnaire (Christie and Geis, 1970), we also included several of our own items designed to make the personality manipulation appear more valid to participants. Because we included our own items and included neither of the actual inventories in its entirety, the questionnaire's value as an actual personality inventory is dubious.

After they completed the personality questionnaire, the experimenter randomly paired participants to play the prisoners' dilemma shown in Fig. 1. The experimenter gave participants a written description of the game and its payoffs, which included Fig. 1 and a detailed description of the payoffs for each combination of choices. The experimenter read the instructions aloud and ensured that participants understood the game. There was no context given for the game: it was labeled simply "the game" and the moves were called "Choice A" and "Choice B".

At this point, the procedure varied by condition. Participants in the *control* condition ($n = 15$) completed the dependent measures. Participants in the *positive-personality* and *negative-personality* conditions (each $n = 15$) were introduced to the person perception task.

2.2.1. Personality manipulation

Participants in the positive- and negative-personality conditions were told they would complete a person perception task designed to study the role of people's impressions of each other in inter dependent decisions such as the one they were about to make. Participants were told they would be assigned to one of two roles, a *target* or a *perceiver*, and that one person from each pair would be assigned to each role. Before making a decision in the game, the perceiver was to form an impression of the target by reading the target's responses to the personality questionnaire. Targets would complete unrelated questionnaires while the perceiver was forming his or her impression.

Participants were randomly assigned and escorted to private rooms where the experimenter had placed instructions designating their role assignment. Inside each room were

instructions assigning participants to the role of perceiver. All participants thus thought that they were going to read the personality questionnaire that had been completed by their partner, whom they believed had been assigned to the role of target.

Participants read one of two versions of the questionnaire that in fact had been completed in advance by the experimenter. Participants in the positive-personality condition read a questionnaire that was engineered to represent the other player in a positive light: *Me* was circled for each of the eight positive statements, *Not Me* was circled for 10 of the 12 negative statements, and *Neutral* was circled for two negative statements.[3] Participants in the negative-personality condition read a questionnaire that was engineered to represent the other player in a negative light: *Me* was circled for 10 of the 12 negative statements, *Not Me* was circled for the positive statements, and *Neutral* was circled for two negative statements. These questionnaires were designed to give participants a strong positive or strong negative impression of their partner.

2.2.2. Dependent measures

Participants in the positive- and negative-personality conditions completed the dependent measures after reading the personality questionnaires; participants in the control condition did so immediately after reading the game instructions. For each of the four possible outcomes of the game, participants indicated how much they liked the outcome by circling a number on scale ranging from *not at all* (1) to *very much* (9). Participants also estimated how much they thought their partner liked each of the four possible outcomes by circling numbers on four identical scales. Half of the participants indicated their own preferences first and half predicted their partner's preferences first. Finally, participants predicted whether the other player would select Choice A or Choice B and then indicated their own choice.

Following completion of the dependent measures, participants were probed for suspicion regarding the personality manipulation. None suspected the personality questionnaire was not genuine. We then informed participants in the positive- and negative-personality conditions that in fact, we had engineered the personality questionnaires. Several participants expressed surprise. We told participants in all three conditions that because we had used deception, they would not actually play the game. Instead, they would each be paid US$ 5.

We thoroughly debriefed participants regarding the reasons for our use of deception in our experiment. We explained that our manipulation provided a strong manipulation of participants' impressions of their partners and that we were interested in the impact of this manipulation on their preferences and decisions in the game. We also asked them not to tell their peers about our use of deception. We return to the issue of deception in Section 4.

3. Results

We present our analysis in three sections. First, we examine whether the personality manipulation was successful by testing the impact of the manipulation on participants'

[3] We judged statements 1, 2, 3, 4, 11, 12, 13, and 20 to be positive and the others to be negative.

prediction of the other player's preferences and move. Second, we examine whether participants chose rationally, given their stated preferences and their estimate of the other player's move (the rational-choice hypothesis). Third, we test for contingent social utility by examining whether the personality manipulation affected participants' stated preferences (the contingent-utility hypothesis). We also investigate whether participants' stated preferences are more contingent on their prediction of their partner's preferences or on their prediction of their partner's move (the motives-versus-moves hypothesis). [4]

Throughout the presentation of our results, we use the language of the prisoners' dilemma rather than the abstract language of the game we showed participants. For example, we refer to "cooperation" and "defection" rather than Choice A and Choice B from Fig. 1.

3.1. Manipulation checks

To examine whether the personality manipulation affected participants' estimate of the other player's preferences we created a variable called OTHER'S PREF by subtracting participants' estimate of how much the other player liked the "temptation" outcome (i.e. cooperation by the participant and defection by the other player) from participants' estimate of how much the other player liked the mutual-cooperation outcome ($M = 0.24$, S.D. $=$ 4.23). Positive values could be as high as eight and indicated that participants thought the other player preferred to match cooperation with cooperation; negative values could be as low as -8 and indicated that participants thought the other player preferred to defect in response to cooperation.

Notice that OTHER'S PREF ignores participants' estimate of the other player's preferences in response to defection. We do this for both theoretical and empirical reasons. The theories of contingent social utility described earlier either are silent on this question or predict that players will prefer to defect in response to defection. Moreover, of our 45 participants, only eight indicated they liked the "sucker" outcome (i.e. cooperation by the participant and defection by the other player) more than mutual defection, and only seven estimated that the other player preferred the sucker outcome (i.e. defection by the participant and cooperation by the other player) over mutual defection. OTHER'S PREF is thus a simple summary of the interesting part of participants' prediction of the other player's preferences in the prisoners' dilemma.

The mean of OTHER'S PREF by experimental condition is displayed in Table 1. An analysis of variance (ANOVA) revealed that our manipulation had an effect ($F(2, 42) =$ 15.97, $P < 0.001$). Planned comparisons indicated that OTHER'S PREF was significantly higher in the positive-personality condition than in both the negative-personality condition ($t = 5.65$, $P < 0.01$) and the control condition ($t = 3.02$, $P < 0.01$), and significantly lower in the negative-personality condition than in the control condition ($t = 2.63$, $P <$ 0.025).

The personality manipulation similarly affected participants' prediction of the other player's move. We defined OTHER'S MOVE to equal to 1 if participants predicted the

[4] We found no effects of the order in which participants estimated their own and the other player's preferences in any of our analyses so they are not discussed hereafter.

Table 1
Participant's estimate of other player's preferences (OTHER'S PREF) and the percentage of participants who predict the other player would choose to cooperate (OTHER'S MOVE), by experimental condition

	Condition		
	Negative-personality	Control	Positive-personality
OTHER'S PREF[a]	−3.20	0.40	3.53
OTHER'S MOVE (percent)	40	87	100

[a] OTHER'S PREF was created by subtracting participants' estimate of how much the other player liked the temptation outcome (i.e. cooperation by the participant and defection by the other player) from participants' estimate of how much the other player liked the mutual-cooperation outcome ($M = 0.24$, S.D. $= 4.23$). It therefore ranges from -8 to $+8$.

other player would cooperate and 0 otherwise ($M = 0.76$, S.D. $= 0.43$). The percentage of participants in each condition that predicted the other player would cooperate is displayed in Table 1. A logistic regression estimating OTHER'S MOVE from a constant and two dummy variables for the positive- and negative-personality conditions correctly predicted other's move 82% of the time ($\chi^2(N = 45, \text{d.f.} = 2) = 18.08$, $P < 0.01$). More participants in both the control and positive-personality conditions predicted that the other player would cooperate than in the negative-personality condition ($P < 0.01$ in both cases). There was no significant difference in OTHER'S MOVE between the control and positive-personality conditions.

3.2. Rational choice

We next examined the rational-choice hypothesis: people choose optimally, given their stated preferences and their prediction of the other player's choice. The number of participants who chose to cooperate or defect is presented in Table 2 by whether they indicated a preference to cooperate, a preference to defect, or indifference (given their prediction of the other player's choice). Excluding those who were indifferent, 74% of participants chose optimally, cooperating or defecting when their stated preferences and their prediction of the other player's move made it rational to do so. Had we assumed that people's preferences were based strictly on their monetary payoffs and hence predicted that no one would cooperate, we would have correctly predicted only 23% of participants' moves. The difference between the percentage of choices correctly predicted based on stated preferences and the

Table 2
Participants' chosen move by their stated preferences, given their prediction of the other player's move

Chosen move	Stated preference given prediction of other player's move		
	Defect	Indifferent	Cooperate
Cooperate	8	4	22
Defect	7	2	2

percentage of choices correctly predicted based on monetary preferences is statistically significant, $P < 0.001$.[5]

To deepen this analysis of rational choice, we ran three logistic regressions. We defined OWN MOVE as a binary variable equal to 1 if the participant chose cooperation and 0 otherwise. We then conducted a logistic regression estimating OWN MOVE from a variable called SIGN PREF, which indicated the direction of participants' stated preferences. That is, given their prediction of the other player's move, SIGN PREF equals +1 if a participant states a preference for cooperation, 0 if a participant is indifferent, and −1 if a participant states a preference for defection. This logistic regression did not contain a constant term, so that we could assess the pure effect of the sign of participants' preferences on their move. This regression correctly predicted OWN MOVE 73% of the time and SIGN PREF was statistically significant ($\beta = 1.06$, $P < 0.005$)

Because the sign of participants' preferences is not perfectly consistent with their moves (only 74% chose rationally), we next examined whether the *strength* of participants' preferences adds explanatory power. We defined STRENGTH PREF as the magnitude of participants' preference for cooperation, given their prediction of the other player's move. If a participant predicted the other player would cooperate, STRENGTH PREF equals that participant's rated liking for mutual cooperation minus her stated liking for the temptation outcome. If a participant predicted the other player would defect, STRENGTH PREF equals that participant's stated liking for the sucker payoff minus her stated liking for mutual defection. We then ran a second logistic regression predicting OWN MOVE from both SIGN PREF and STRENGTH PREF, again without a constant. This regression correctly predicted OWN MOVE 72% of the time but only SIGN PREF was significant ($\beta = 0.99$, $P < 0.09$); STRENGTH PREF was not ($\beta = 0.02$, $P = 0.86$). Thus, although the sign of participants' preferences does not perfectly explain their chosen action, inconsistencies between stated preferences and chosen action are no less likely when stated preferences are far from indifference than when they are near.

There is a high rate of cooperation in our sample, as seen in Table 2. For example, of the 15 participants who indicated that they preferred to defect given their prediction of the other player's move, eight chose to cooperate. To account for this high rate of cooperation, we re-ran the first logistic regression above, but this time with a constant. This third regression correctly predicted OWN MOVE 76% of the time and SIGN PREF was statistically significant ($\beta = 1.10$, $P < 0.025$), as was the constant ($\beta = 1.56$, $P < 0.005$). This finding of a significant constant term, like the fact that 26% of participants did not choose optimally, casts some doubt on either the precision of our stated preference variable as a measure of utility or on the assumption that people choose optimally given their preferences. Nonetheless, these findings are generally supportive of our rational-choice hypothesis: people choose optimally given their stated preferences and their prediction of the other player's move.

[5] The specific test is a binomial comparison of the participants whose move was correctly predicted by their stated preferences but not by pure monetary preferences versus the participants whose move was correctly predicted by their monetary preferences but not by their stated preferences, i.e. the 22 participants who indicated they preferred to cooperate and did so versus the two participants who indicated they preferred to cooperate but actually defected.

Table 3
Participant's stated preference for cooperation (OWN PREF) by experimental condition

	Condition		
	Negative-personality	Control	Positive-personality
OWN PREF[a]	0.60	1.87	3.80

[a] OWN PREF was created by subtracting participants' rating of how much they liked their temptation outcome (i.e. defection by the participant and cooperation by the other player) from their rating of how much they liked the mutual-cooperation outcome ($M = 2.09$, S.D. $= 3.53$). It therefore ranges from -8 to $+8$.

3.3. Is social utility contingent? And, if so, on what?

To summarize participants' stated preferences, we created the variable OWN PREF by subtracting participants' rating of how much they liked their temptation outcome from their rating of how much they liked the mutual-cooperation outcome ($M = 2.09$, S.D. $= 3.53$). Recall that OTHER'S PREF measured participants' estimate of the other player's preference for cooperation in response to cooperation; OWN PREF does the same for participants' own preferences. The mean of OWN PREF by experimental condition is displayed in Table 3. As hypothesized, participants' preferences were contingent, as indicated by a one-way ANOVA on OWN PREF by experimental condition ($F(2, 42) = 3.48$, $P < 0.05$). Participants in the positive-personality condition preferred mutual cooperation more than did participants in the control condition ($P = 0.12$) or participants in the negative-personality condition ($P < 0.05$). There was no significant difference in OWN PREF between the negative-personality and control conditions ($P = 0.3$). This finding is consistent with our contingent-utility hypothesis: people's preferences are contingent on their impression of the other player.

Social utility may be contingent, but contingent on what? As described in Section 1, two possibilities for the source of contingency in games such as the prisoners' dilemma: the *moves hypothesis*, formalized by Rabin (1993), is that social utility is contingent on people's prediction of the other player's action; the *motives hypothesis*, formalized by Levine (1998), is that social utility is contingent on people's estimate of the other player's motives. Because our participants both estimated the other player's preferences and predicted the other player's move, our data allow us to examine whether participants' preferences were contingent on motives or on moves.

To distinguish between these two sources of contingency, we conducted a linear regression predicting OWN PREF from two regressors: OTHER'S MOVE (the moves hypothesis) and OTHER'S PREF (the motives hypothesis). The regression was significant ($R^2 = 0.24$, $P < 0.01$). The coefficient on OTHER'S PREF was significant ($\beta = 0.34$, $P < 0.01$), but the coefficient on OTHER'S MOVE was not ($\beta = 1.04$, $t < 1$), suggesting that social utility is contingent on motives more than on moves.

Because OTHER'S MOVE and OTHER'S PREF are highly correlated ($r = 0.48$, $P < 0.001$), we also performed a more conservative test of the source of contingency. We first regressed OWN PREF on OTHER'S MOVE ($\beta = 2.64$, $P < 0.05$), saving the residuals. OTHER'S MOVE was thus able to absorb as much variance as possible in OWN PREF (whether through its direct effect or through correlation with omitted variables, possibly including OTHER'S PREF). We next regressed those residuals on OTHER'S PREF, which

was significant ($\beta = 0.27$, $P < 0.025$). Thus, participants' estimate of the other player's preferences provide explanatory power beyond that provided by participants' prediction of the other player's move.[6] These regression results suggest a tentative answer to the motives-versus-moves hypothesis: people's preferences are contingent on their estimate of the other player's preferences more than on their prediction of the other player's move.

The combination of findings reported in this sub-section-the ANOVA showing that OWN PREF varies by experimental condition and the regressions showing that OWN PREF varies with OTHER'S PREF more than with OTHER'S MOVE-leaves one question unanswered: is OTHER'S PREF simply proxying for some other factor that varies by experimental condition? To answer this question, we conducted a linear regression of OWN PREF on OTHER'S PREF and two dummies for the experimental conditions. The regression was significant ($R^2 = 0.24$, $P < 0.015$). The coefficient on OTHER'S PREF was significant ($\beta = 0.34$, $P < 0.035$), an F-test for the joint significance of the dummies was not significant ($F < 1$). This regression suggests that the key feature of our experimental conditions is their effect on participants' estimate of the other player's preferences.[7] That is, the effect of experimental condition on OWN PREF was mediated by the effect of experimental condition on OTHER'S PREF.

4. Discussion

We examined people's stated preferences for outcomes in a prisoners' dilemma when they had either a favorable or unfavorable impression of the other player. We found that most participants chose optimally, given their stated preference and their prediction of the other player's move. More important, our results support the central assumption of the Rabin (1993), Levine (1998), and Sally (1999) theories that social utility may be contingent: Participants who had a favorable impression of the other player tended to prefer mutual cooperation more than participants who had an unfavorable impression of the other player. Furthermore, our results suggest that participants' preferences were contingent on their estimate of the other player's preferences more than they were contingent on their prediction of the other player's move (or any other unmeasured aspect of our experimental conditions). This latter result is directly supportive of Levine's theory and broadly supportive of Rabin's and Sally's theories.

Two aspects of our data relate to earlier work in social psychology. First, the strong correlation between people's own preferences and their perceptions of the other player's preferences is reminiscent of the false consensus effect, or the tendency for people's own perceptions and behaviors to be positively correlated with their estimates of the commonness of those perceptions and behaviors among their peers (Kelley and Stahelski, 1970; Ross et al., 1977; Marks and Miller, 1987). Dawes (1989) notes that if people are Bayesian

[6] We also estimated the reverse pair of regressions. We first regressed OWN PREF on OTHER'S PREF ($\beta = 0.40$, $P < 0.001$) and saved the residuals. We then regressed those residuals on OTHER'S MOVE, which was not significant ($\beta = 0.81$, $P = 0.46$). Again, social utility appears to be contingent on motives more than on moves.

[7] We also regressed OWN PREF on OTHER'S PREF, OTHER'S MOVE, and dummies for the experimental condition. The results were very similar: OTHER'S PREF was significant and nothing else was.

then their belief about the behavior of the population of others will (quite correctly) depend on their own behavior, because their own behavior is the only signal they have about the population.[8] But neither the original Ross et. al. interpretation nor the subsequent Dawes interpretation provides a natural explanation for our finding that people's preferences differ across experimental conditions. We found that participants who were randomly given a positive impression of the other player were more likely to prefer mutual cooperation than were participants who were randomly given a negative impression of the other player. Although the false consensus effect might explain a positive correlation between participants' own preferences and their estimates of the other player's preferences in the control condition, it offers no explanation for the observed variation in participants' preferences across experimental conditions.

Second, a substantial body of evidence suggests that some people are habitual cooperators and others are habitual defectors (Kelley and Stahelski, 1970; Kuhlman and Marshello, 1975; Komorita and Parks, 1995). Because participants did not play our game repeatedly, and because our personality questionnaire was bogus, our data do not allow us to examine whether there were stable individual differences in social utility among participants in our study. But it is possible that both the distributional view and the contingent view of social utility are partially correct.

Three aspects of our experimental manipulation may cause some concern. In each case, however, we argue that the concern, although important in its own right, is unlikely to overturn our main finding that social utility is contingent. First, because participants knew that their stated preferences and moves were not completely anonymous and would be seen by the experimenter, they may have experienced evaluative concerns that might have affected their responses. For example, the relatively high rate of cooperation in our sample may be due partly to participants' desire to convey a positive impression to the experimenter and to the other player. It should be noted, though, that despite such concerns, a third of our sample (15 of 45) stated that they preferred to defect and almost one-quarter (11 of 45) chose to defect. But even if self-presentational concerns increased the overall rate of stated preferences for cooperation, they do not provide an alternative interpretation of the differences across conditions in participants' stated preferences for cooperation. Social utility is still contingent.

Second, the extreme personality questionnaire responses that we engineered in the negative-personality condition may not have been representative of actual responses in a given population. In particular, given the pervasiveness of self-serving biases (Babcock and Loewenstein, 1997) and people's tendencies to see themselves in a favorable light, the responses of people in the negative-personality condition may have appeared to participants as highly unlikely and extreme. But the fact that our manipulations were not representative of people's statements about their own personalities in everyday interactions again does not affect our key finding that social utility was contingent on those (fictional) statements.

Third, if participants became aware of our use of deception, they may have become suspicious of our manipulations and so their responses may not have reflected their preferences or actions in the real world. Recall, however, that no participant spontaneously mentioned

[8] Krueger and Clement (1994) have shown that people overweight the signal of their own behavior even when provided with information about the population.

being suspicious about our manipulation. A similar concern is that after participants learned about the deception, they may have told their peers who may have later participated in our study (Lichtenstein, 1970; Aronson, 1966). But issues of this kind arise in any experiment in which participants are involved over time rather than all at once, and we took standard precautions to avoid such difficulties (such as conducting our study within a 2 week period and asking participants not to inform their peers about our use of deception). Furthermore, this concern again does not provide a ready interpretation of our findings.

Our use of deception also raises broader concerns about the treatment of research participants. Whether to use deception or not hinges on the researchers' assessment of whether the costs of deception outweigh the potential benefits to be gleaned from the research. We believe such an assessment favored the use of deception in this particular case because it allowed a clean test of an important question in behavioral science. Had we found that even our extreme personality manipulation did not cause social utility to be contingent on impressions of the other player, we would have viewed theories that begin from this assumption with substantial skepticism. Of course, given our findings, it now becomes of interest to know whether more realistic manipulations can cause similar effects.

In sum, we hope to have contributed to a new strand of research in behavioral game theory that we believe will become quite important: the application of tenets from interpersonal perception and other parts of social psychology to the settings analyzed in experimental game theory. In addition to our focus on social utility in behavioral game theory (see also Camerer, 1997), there is also work being done on self-serving biases (Babcock and Loewenstein, 1997), egocentrism (Van Boven et al., 2000; Van Boven et al., 2000), and attributions in games (Durell, 1999; Weber et al., 2000). We expect that just as the incorporation of heuristics and biases has been productive for behavioral decision theory, the incorporation of social psychological research will be productive for behavioral game theory.

Acknowledgements

Author order is alphabetical. Max Bazerman, Iris Bohnet, Colin Camerer, Robyn Dawes, Tom Gilovich, Keith Murnighan, Tom Ross, and two anonymous reviewers made helpful comments on earlier versions of this paper. Cornell University's Johnson School of Management, MIT's Sloan School of Management, and the NSF (Grant SBR-9809107) provided financial support.

Appendix A. Personality questionnaire

This questionnaire is designed to tell us what kind of person you are. Your responses are very important for our data analysis. Please be as honest as possible.

Instructions: for each statement, please indicate how much the statement is characteristic of you by circling one of the following:

- Me
- Neutral
- Not Me

Rating			Statement
Me	Neutral	Not me	I am sincere and trustworthy; I will not lie, for whatever ends
Me	Neutral	Not me	I pride myself on being highly principled; I am willing to stand by those principles no matter what the cost
Me	Neutral	Not me	My sense of humor is one of my biggest assets
Me	Neutral	Not me	I have above-average empathy for the views and feelings of others
Me	Neutral	Not me	I like power; I want it for myself, to do with what I want; In situations where I must share power, I strive to increase by power base, and lessen my co-power holder's power base
Me	Neutral	Not me	I enjoy trying to persuade others to my point of view
Me	Neutral	Not me	I feel if I am too honest and trustworthy, most people will take advantage of me
Me	Neutral	Not me	To persuade others, I prefer to use fear rather than trust
Me	Neutral	Not me	I try not to be predictable because then I can be easily manipulated
Me	Neutral	Not me	love to be the aggressor; I believe I have to take the initiative if I want to accomplish my goals
Me	Neutral	Not me	I believe honesty and openness are essential for maintaining good relationships
Me	Neutral	Not me	In a negotiation, I believe the best outcome is one that is fair for all parties
Me	Neutral	Not me	I believe one can achieve the best results in life by cooperating with others
Me	Neutral	Not me	I believe one can achieve the best results in life by competing with others
Me	Neutral	Not me	I believe principles are fine for some people, but sometimes they have to be sacrificed to achieve one's goals
Me	Neutral	Not me	In negotiations, I try to exploit my opponent's weaknesses
Me	Neutral	Not me	I believe that imposing personal discomfort is not too high a price to pay for success in negotiation
Me	Neutral	Not me	I believe there is nothing wrong with lying in a competitive situation, as long as I don't get caught
Me	Neutral	Not me	I believe sometimes I must sacrifice the welfare of others for my own benefit
Me	Neutral	Not me	I believe that the dignity and welfare of others is the most important concern for any society

References

Aronson, E., 1966. Avoidance of inter-subject communication. Psychol. Reports 19, 238.
Babcock, L., Loewenstein, G., 1997. Explaining bargaining impasse: the role of self-serving biases. J. Econ. Persp. 11, 109–126.
Bazerman, M., 1998. Judgment in managerial decision making. Wiley, New York.
Bolton, G., 1991. A comparative model of bargaining: theory and evidence. Am. Econ. Rev. 81, 1096–1136.
Blount, S., 1995. When social outcomes are not fair: the effect of causal attributions on preferences. Org. Behav. Human Dec. Proc. 63, 131–144.
Camerer, C., 1990. Behavioral game theory. In: Hogarth, R. (Ed.), Insights in Decision Making. University of Chicago Press, Chicago.
Camerer, C., 1995. Individual decision making. In: Kagel, J., Roth, A. (Eds.), The Handbook of Experimental Economics. Princeton University Press, Princeton, Chapter 8.
Camerer, C., 1997. Progress in behavioral game theory. J. Econ. Persp. 11, 167–188.
Camerer, C., Thaler, R., 1995. Ultimatums, dictators, and manners. J. Econ. Persp. 9, 209–219.
Cameron, L., 1999. Raising the stakes in the ultimatum game: experimental evidence from Indonesia. Econ. Inquiry 37, 47–59.
Charness, G., 1996. Attribution and reciprocity in a simulated labor market: an experimental investigation. Department of Economics, Berkeley, unpublished.
Christie, R., Geis, F.L., 1970. Studies in Machiavellianism. Academic Press, New York.
Dawes, R.M., 1988. Rational Choice in an Uncertain World. Harcourt Brace Jovanovich, San Diego.
Dawes, R.M., 1989. Statistical criteria for establishing a truly false consensus effect. J. Exp. Social Psychol. 25, 1–17.
Dawes, R.M., McTavish, J., Shaklee, H., 1977. Behavior, communication, and assumptions about other people's behavior in a common dilemma situation. J. Personality Social Psychol. 35, 1–11.
Durell, A., 1999. Attribution in performance evaluation. Department of Economics, Dartmouth College, unpublished.
Fehr, E., Falk, A., 1999. Wage rigidity in a competitive incomplete contract market. J. Political Econ. 107, 106–134.
Fehr, E., Gächter, S., Kirchsteiger, G., 1997. Reciprocity as a contract enforcement device. Econometrica 65, 833–860.
Fehr, E., Kirchsteiger, G., Riedl, A., 1993. Does fairness prevent market clearing? An experimental investigation. Q. J. Econ. 108, 437–460.
Forsyth, D.R., 1980. A taxonomy of ethical ideologies. J. Personality Social Psychol. 39, 175–184.
Gilbert, D.T., 1999. Ordinary personology. In: Gilbert, D.T., Fiske, S.T., Gardiner, L. (Eds.), The Handbook of Social Psychology, 4th Edition, Vol. 2. Mcgraw-Hill, Boston, MA, pp. 89–150.
Jones, E.E., 1990. Interpersonal Perception. Macmillan, New York.
Kelley, H.H., Stahelski, A., 1970. Social interaction basis of cooperators' and competitors' beliefs about others. J. Personality Social Psychol. 16, 66–91.
Kelley, H.H., Thibaut, J.W., 1978. Interpersonal Relations: A Theory of Interdependence. Wiley, New York.
Komorita, S.S., Parks, C.D., 1995. Interpersonal relations: mixed-motive interaction. Annu. Rev. Psychol. 46, 495–530.
Krueger, J., Clement, R.H., 1994. The truly false consensus effect: an ineradicable and egocentric bias in social perception. J. Personality Social Psychol. 67, 596–610.
Kuhlman, D.M., Marshello, A., 1975. Individual differences in game motivation as moderators of pre-programmed strategic effects in prisoner's dilemma. J. Personality Social Psychol. 32, 922–931.
Ledyard, J., 1995. Public goods: a survey of experimental research. In: Kagel, J., Roth, A. (Eds.), Handbook of Experimental Economics. Princeton University Press, Princeton, NJ.
Levine, D., 1998. Modeling altruism and spitefulness in experiments. Rev. Econ. Dyn. 1, 593–622.
Lichtenstein, E., 1970. Please do not talk to anyone about this experiment: disclosure of deception by debriefed subjects. Psychol. Reports 26, 459–479.
Loewenstein, G., Thompson, L., Bazerman, M., 1989. Social utility and decision making in interpersonal contexts. J. Personality Social Psychol. 57, 426–441.
Marks, G., Miller, N., 1987. Ten years of research on the false-consensus effect: an empirical and theoretical review. Psychol. Bull. 102, 72–90.

Messick, D., Sentis, K., 1985. Estimating social and non-social utility functions from ordinal data. Eur. J. Social Psychol. 15, 389–399.
Neale, M., Bazerman, M., 1991. Cognition and Rationality in Negotiation. Free Press, New York.
Pillutla, M., Murnighan, J.K., 1996. Unfairness, anger, and spite: emotional rejections of ultimatum offers. Org. Behav. Human Dec. Proc. 68, 208–224.
Rabin, M., 1993. Incorporating fairness into game theory and economics. Am. Econ. Rev. 83, 1408–1418.
Rabin, M., 1998. Psychology and Economics. J. Econ. Lit. 36, 11–46.
Ross, L., Greene, D., House, P., 1977. The false consensus effect: an ego-centric bias in social perception and attribution processes. J. Exp. Social Psychol. 13, 279–301.
Sally, D., 1999. A sympathetic look at the prisoners' dilemma. Johnson Graduate School of Management, Cornell University, unpublished.
Slonim, R., Roth, A.E., 1998. Learning in high stakes ultimatum games: an experiment in the Slovak Republic. Econometrica 66, 569–596.
Van Boven, L., Dunning, D., Loewenstein, G., 2000. Egocentric empathy gaps between owners and buyers: misperceptions of the endowment effect. J. Personality Social Psychol. 79, 66–76.
Van Boven, L., Loewenstein, G., Dunning, D., 2000. Egocentric perceptions of other people's tastes: buyers and sellers' misperceptions of the endowment effect. University of British Columbia, unpublished.
Weber, R., Rottenstreich, Y., Camerer, C., Knez, M., 2000. The illusion of leadership misattribution of course co-ordination games. Org. Sci., in press
Wyer, R.S., 1969. Prediction of behavior in two-person games. J. Personality Social Psychol. 13, 222–238.