Contingent Social Utility in the Prisoners' Dilemma*

Robert Gibbons
MIT's Sloan School and NBER
rgibbons@mit.edu

Leaf Van Boven
Department of Psychology, Cornell University
ldv2@cornell.edu

First draft: November 16, 1997

This version: August 4, 1999

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 (rational choice), that participants' preferences would be contingent on their perception of the other player (contingent utility), and that 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 (motives versus moves). Our evidence supported all three predictions.

*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.

Contingent Social Utility in the Prisoners' Dilemma

R. Gibbons and L. Van Boven

Behavioral decision theory has made important progress by incorporating heuristics and biases that people use in everyday judgment into formal models of single-person decision making (Bazerman 1998; Camerer, 1995; Dawes, 1988). 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 benefit by borrowing other ideas from psychology, such as those on interpersonal perception (Jones, 1990). 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). Research on ultimatum games, in contrast, 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 might 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.

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: 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 some public-good games, the fraction of participants who seem altruistic is above 50%, whereas in some ultimatum games the

---

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 I propose to divide an initial pie of $10 into

fraction who seem spiteful 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) explores some of these implications for ultimatum games and finds that they are not terribly consistent with the data. In short, while 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 x 2 game as depending in part 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 (1993) model, then, people's social utility is contingent on their prediction of the other player's action. We label this 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. For example, player 1 cares positively about player 2's payoff if 1 believes that 2 cares positively about 1's payoffs. Conversely, player 1 cares negatively about player 2's payoff if 1 believes that 2 cares negatively about 1's payoffs.

$7 for me and $3 for you, which you reject, only to propose that we divide the subsequent pie of $5 into $2.50 for each of us.

In Levine's (1998) model, then, people's social utility is contingent on their estimate of the other player's preferences, or motives. We label this the *motives hypothesis*.[2]

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 many fields.

2 Rabin (1998: 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.

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. For a rational player, moves follow in part from motives, so the two should be positively correlated. But to the extent that the correlation is not perfect, we can (and do) compare the predictive power of these two hypotheses.

Empirical Research Related to Contingent Social Utility

Of course, theories of contingent social utility are not grounded solely in intuition but rather are based on a large body of evidence suggesting that social utility is indeed contingent. Early work in social psychology, for example, found that people's decision to cooperate or defect in a Prisoners' Dilemma was based in part on their prediction of whether the other player would cooperate or defect (Dawes, McTavish, and Shaklee, 1977; Kelley and Stahelski, 1970). One interpretation of these findings is that the positive correlation between people's own action and their prediction of the other player's action reflects 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 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.

Finally, Fehr and colleagues have found evidence of reciprocity in experimental economic environments (Fehr, Kirchsteiger, and Riedl, 1993; Fehr, Gächter, and Kirchsteiger, 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 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 perception of the other.

The closest predecessor to our study is Loewenstein, Thompson, and Bazerman (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 dislike payoff disparities (disliking unfavorable disparities more than favorable disparities), but that this was true only for positive, non-business relationships. Their findings thus 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.

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. We address these issues in the Discussion.

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 three 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*); finally, we predicted that 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*).

Method

Participants

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

Procedure

The experimenter asked participants when they arrived at the lab to complete a "personality questionnaire," explaining that their responses would be used to analyze their behavior in the experiment. The personality questionnaire 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*. (The questionnaire is reproduced in the Appendix.)

Although the personality questionnaire included some items culled from actual psychological inventories, notably the Ethics Position Questionnaire (Forsyth, 1980) and the Machiavellianism questionnaire (Chrisie and Geis, 1970), we also included several of our own items designed to give the personality manipulation discussed below more punch. 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 Figure 1. The experimenter gave participants a written description of the game and its payoffs, which included Figure 1 and a detailed description of the payoffs for each combination of choices. The experimenter read the instructions out loud 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. For participants in the *control* condition (n = 15), the experimenter asked participants to complete the dependent measures described below. For those in *positive personality* and *negative personality* conditions (each n = 15), the experimenter introduced the participants to the "person perception task."

Personality manipulation. Participants in the positive- and negative-personality conditions were told they would complete a "person perception task" designed to study how people form impressions of each other in interdependent 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 personality questionnaire. Targets would complete unrelated questionnaires while the perceiver was forming his or her impression.

The experimenter said participants would be randomly assigned to different rooms where he had placed instructions designating their role assignment. Pairs of participants flipped a coin to determine which room they entered. Inside both rooms were instructions assigning participants to the role of perceiver. All participants thus thought that they were going to read their partner's personality questionnaire and that their partner had been assigned to the role of target.

Participants received one of two questionnaires that had been completed in advance. Participants in the positive-personality condition received a questionnaire that had been engineered to represent the other player in a positive light: *Me* was circled for each of the 8 positive statements, *Not Me* was circled for 10 of the 12 negative statements, and *Neutral* was

circled for two negative statements.[3] The reverse was true in the negative-personality version: *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.

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 a scale ranging from *not at all* (1) to *very much* (9). Participants also predicted how

3 We judged statements 1, 2, 3, 4, 11, 12, 13, and 20 to be positive and the others to be negative. See the Appendix for the full questionnaire.

much they thought their partner liked each of the four possible outcomes by circling numbers on an identical scale. 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 generated by the other player. 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 $5.

We then 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 this issue in the Discussion.

Results

We present our analysis in three sections. First, we examine whether our personality manipulation was successful by testing the impact of the manipulation on participants' 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 conduct a preliminary comparison of the motives and moves hypotheses by examining whether participants' stated preferences are contingent on their prediction of their partner's preferences or their partner's move (the motives-versus-moves hypothesis).[4]

---

4 There were no order effects in any of our analyses so they are not discussed hereafter.

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 Figure 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 = .24, sd = 4.23). Positive values could be as high as 8 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 8 indicated they liked their sucker outcome (*i.e.*, cooperation by the participant and defection by the other player) more than mutual defection, and only 7 estimated that the other player preferred their 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.

Table 1(a) shows the mean of OTHER'S PREF by experimental condition. An analysis of variance (ANOVA) shows that our manipulation had an effect ($F(2,42) = 15.97$, $p < .001$). Furthermore, planned comparisons indicated that OTHER'S PREF was significantly higher in the positive-personality condition than in both the negative-personality condition ($t(42) = 5.65$, $p < .01$) and the control condition ($t(42) = 3.02$, $p < .01$), and significantly lower in the negative-personality condition than in the control condition ($t(42) = 2.63$, $p < .025$).

The personality manipulation similarly affected participants' prediction of the other player's move. We defined OTHER'S MOVE to equal 1 if participants predicted the other player would cooperate and 0 otherwise (M = .76, sd = .43). 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 (χ^2(N = 45, df = 2) = 18.08, p < .01). Table 1(b) shows that in both the control and positive-personality conditions more participants predicted that the other player would cooperate than in the negative-personality condition (p < .01 in both cases). There was no significant difference between the control and positive-personality conditions in the percentage of participants who predicted the other player would cooperate.

Rational Choice

We next examined our first hypothesis—that people would choose optimally, given their stated preferences and their prediction of the other player's choice. Table 2 presents the number of people who chose to cooperate by whether they indicated a preference to cooperate, a preference to defect, or indifference (given their prediction of the other player's choice). Excluding people who were indifferent, 74% of participants chose optimally: they cooperated or defected 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 everyone would defect, 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 percentage of choices correctly predicted based on monetary preferences is statistically significant, p < .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 estimated a logistic regression predicting OWN MOVE from a variable called SIGN PREF equal to the sign of the participants' stated preference for cooperation over

---

5 The specific test is a binomial comparison of the 22 participants whose move was correctly predicted by their stated preferences but not predicted by pure monetary preferences—participants who indicated they preferred to cooperate and did so—and the 2 participants who indicated they preferred to cooperate but actually defected.

defection, given their prediction of the other player's move.[6] 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 action, just as in Table 2. This regression correctly predicted OWN MOVE 73% of the time and SIGN PREF was statistically significant ($\beta = 1.06$, $p < .005$)

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 ran a second logistic regression predicting OWN MOVE from both SIGN PREF and STRENGTH PREF, again without a constant.[7] This regression correctly predicted OWN MOVE 72% of the time but only SIGN PREF was significant ($\beta = .99$, $p < .9$); STRENGTH PREF was not ($\beta = .02$, $p = .86$). Thus, although the sign of participants' preferences does not perfectly explain their chosen action, inconsistencies between stated preferences and chosen action are no more likely when stated preferences are far from indifference than when they are near.

Table 2 shows that there is a high rate of cooperation in our sample. For example, of the 15 participants who indicated that they preferred to defect given their prediction of the other player's move, 8 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 ($\beta = 1.10$, $p < .025$) as was the constant ($\beta = 1.56$, $p < .005$). This finding of a significant constant term, like the fact that 26% of participants did not choose optimally, casts some doubt on either the

---

6 That is, SIGN PREF equals 1 if the participant stated a preference for cooperation, -1 for defection, and 0 for indifference. To compute SIGN PREF, we first created the variable OWN PREF 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 ($M = 2.09$, $sd = 3.53$). OWN PREF is analogous to OTHER'S PREF in measuring only the preferred response to cooperation. But to examine rational choice by participants, we must also consider their stated preference in response to defection. We therefore also created OWN PREF (DEF) by subtracting participants' rating of how much they liked the mutual-defection outcome from their rating of how much they liked their sucker outcome (*i.e.*, cooperation by the participant and defection by the other player) ($M = -1.76$, $sd = 2.88$). Finally, from OWN PREF and OWN PREF (DEF) we created a third variable, STRENGTH PREF, which equals OWN PREF if the participant predicted that the other player would cooperate but equals OWN PREF (DEF) if the participant predicted that the other player would defect ($M = 1.27$, $sd = 4.24$). That is, STRENGTH PREF = OWN PREF * OTHER'S MOVE + OWN PREF (DEF) * [1 – OTHER'S MOVE], and SIGN PREF then equals the sign of STRENGTH PREF (*i.e.*, +1 for positive values, -1 for negative values, and 0 otherwise).

accuracy 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—that people choose optimally given their stated preferences and their prediction of the other player's move.

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 (*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, sd = 3.53). Recall that OTHER'S PREF measured participants' estimate of the other player's preference for cooperation in response to cooperation;

---

7 See footnote 6 for a definition of STRENGTH PREF.

OWN PREF does the same for participants' own preferences. Table 3(a) shows the mean of OWN PREF by experimental condition. As hypothesized, participants' preferences were contingent, as indicated by an ANOVA on OWN PREF by experimental condition ($F(2, 42) = 3.48$, $p < .05$). Participants in the positive-personality condition preferred mutual cooperation more ($M = 3.80$) than did participants in the control condition ($M = 1.87$, $p = .12$) or participants in the negative-personality condition ($M = .60$, $p < .05$). There was no significant difference in OWN PREF between the negative-personality and control conditions ($p = .3$). This finding is consistent with our contingent-utility hypothesis—that people's preferences are contingent on their impression of the other player.

It is of some interest to know that social utility can be contingent, but it would be of greater interest to know contingent on what. To our knowledge, two possibilities have been proposed for 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 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 estimated 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 ($R^2 = .24$, $p < .01$). The coefficient on OTHER'S PREF was significant ($\beta = .34$, $p < .01$), but the coefficient on OTHER'S MOVE was not ($\beta = 1.04$, $t < 1$). This regression suggests that social utility is contingent on motives rather than moves.

But OTHER'S MOVE and OTHER'S PREF are highly correlated ($r(44) = .48$, $p < .001$), so interpretation of the regression above may be difficult. To perform a more conservative test, we first regressed OWN PREF on OTHER'S MOVE ($\beta = 2.64$, $p < .05$), 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

($\beta = .27$, $p < .025$). Thus, participants' estimate of the other player's preferences provide explanatory power beyond that provided by participants' prediction of the other player's move.[8] These regression results are consistent with our motives-versus-moves hypothesis—that participants' preferences are contingent on their estimate of the other player's preferences, rather 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 rather than with OTHER'S MOVE—leaves one question unanswered: is OTHER'S PREF simply proxying for something else in our experimental conditions? To answer this question, we estimated a linear regression of OWN PREF on OTHER'S PREF and two dummies for the experimental conditions. The regression was significant ($R^2 = .24$, $p < .015$). The coefficient on OTHER'S PREF was significant ($\beta = .34$, $p < .035$), but the coefficients on the dummies were not ($t < .05$ and $t < .6$), and an F-test for the joint significance of the dummies also was not significant ($F(2, 41) = .27$, $p = .77$). This finding is consistent with the idea that the key feature of our experimental conditions is the effect of the condition on participants' estimate of the other player's preferences.[9]

Discussion

We examined the stated preferences for outcomes in a Prisoners' Dilemma of participants who had 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 importantly, 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 were more likely to prefer to cooperate in response to cooperation than were participants who had an unfavorable impression of the other

---

8 We also estimated the reverse pair of regressions. We first regressed OWN PREF on OTHER'S PREF ($\beta = .40$, $p < .001$) and saved the residuals. We then regressed those residuals on OTHER'S MOVE, which was not significant ($\beta = .81$, $p = .46$). Again, social utility appears to be contingent on motives rather than on moves.

9 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.

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 (Kelley and Stahelski, 1970; Marks and Miller, 1985; Ross, Greene, and House, 1977), which suggests that people tend to see their own attitudes, beliefs, and behaviors as disproportionately common among their peers. Dawes (1989) notes that if people are Bayesian 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.[10] 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 randomly selected participants who received a positive personality questionnaire about the other player were more likely to prefer to meet cooperation with cooperation than were participants who received a negative personality questionnaire. Thus, although the false consensus effect might explain a correlation between a player's own preferences and that player's perceptions of the other player's preferences, it offers no explanation for the variation in players' own preferences across experimental conditions.

Second, a substantial body of psychological evidence suggests that some people are habitual cooperators and others are habitual defectors (Komorita and Parks, 1995; Kelley and Stahelski, 1970; Kuhlman and Marshello, 1975). 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, even likely, that both the distributional view and the contingent view of social utility are partially correct, so theorists might incorporate both into descriptive models of social utility.

Three aspects of our experimental manipulation may cause some concern. In each case, however, we argue that the concern, while 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. Indeed, the relatively high rate of cooperation in our sample may be due, in part, 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 difference between 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, our use of deception may have led some participants to 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 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; but see 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 two-week period and asking participants not

---

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

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—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 measurable 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 long-standing tenets from interpersonal perception and other parts of social psychology to the settings analyzed in experimental game theory. An analogous marriage was productive for behavioral decision theory, but the fact that games have more than one person suggests that new ideas from social psychology will be relevant in behavioral game theory. In addition to our focus on social utility (see also Camerer, 1997), there is exciting work being done on self-serving biases (Babcock and Loewenstein, 1997), egocentrism (Van Boven, Dunning, and Loewenstein, 1999), and attributions in games (Durell, 1999; Weber, Rottenstreich, Camerer, and Knez, forth.). We anticipate an active decade of such research.

References

Aronson, E. 1966. “Avoidance of inter-subject communication.” *Psychological Reports, 19*, 238.

Babcock, L. and G. Loewenstein. 1997. “Explaining bargaining impasse: The role of self-serving biases.” *Journal of Economic Perspectives* 11: 109-26.

Bazerman, M. 1998. *Judgment in managerial decision making*. New York: John Wiley & Sons.

Bolton, G. 1991. “A comparative model of bargaining: Theory and evidence.” *American Economic Review* 81: 1096-1136.

Blount, S. 1995. “When social outcomes aren’t fair: The effect of causal attributions on preferences.” *Organizational Behavior and Human Decision Processes* 63: 131-44.

Camerer, Colin. 1990. “Behavioral Game Theory.” In R. Hogarth (ed.), *Insights in Decision Making*. Chicago: University of Chicago Press.

_____. 1995. “Individual Decision Making.” Chapter 8 in J. Kagel and A. Roth (eds.), *The Handbook of Experimental Economics*. Princeton: Princeton University Press.

_____. 1997. “Progress in Behavioral Game Theory.” *Journal of Economic Perspectives* 11: 167-88.

_____ and R. Thaler. 1995. “Ultimatums, Dictators, and Manners.” *Journal of Economic Perspectives* 9: 209-19.

Cameron, L. 1999. “Raising the Stakes in the Ultimatum Game: Experimental Evidence from Indonesia.” *Economic Inquiry* 37: 47-59.

Charness, G. 1996. “Attribution and Reciprocity in a Simulated Labor Market: An Experimental Investigation.” Unpublished manuscript, Department of Economics, Berkeley.

Christie, R., & Geis, F. L. (1970). *Studies in Machiavellianism*. New York: Academic Press.

Dawes, R. M. 1989. “Statistical criteria for establishing a truly false consensus effect.” *Journal of Experimental Social Psychology, 25*, 1-17.

_____. 1988. *Rational choice in an uncertain world*. San Diego: Harcourt Brace Jovanovich.

_____, McTavish, J., and Shaklee, H. 1977. “Behavior, communication, and assumptions about other people’s behavior in a common dilemma situation.” *Journal of Personality and Social Psychology, 35,* 1-11.

Durell, A. 1999. “Attribution in Performance Evaluation.” Unpublished manuscript, Department of Economics, Dartmouth College.

Fehr, E. and A. Falk. 1999. “Wage Rigidity in a Competitive Incomplete Contract Market.” *Journal of Political Economy* 107: 106-34.

_____, S. Gächter, and G. Kirchsteiger. 1997. “Reciprocity as a Contract Enforcement Device.” *Econometrica* 65: 833-60.

_____, G. Kirchsteiger, and A. Riedl. 1993. “Does Fairness Prevent Market Clearing? An Experimental Investigation.” *Quarterly Journal of Economics* 108: 437-60.

Forsyth, D. R. (1980). A taxonomy of ethical ideologies. *Journal of Personality and Social Psychology, 39,* 175-84.

Jones, E.E. 1990. *Interpersonal perception*. New York: Macmillan.

Kelley, H. H., and Stahelski, A. 1970. “Social Interaction Basis of Cooperators’ and Competitors’ Beliefs about Others.” *Journal of Personality and Social Psychology* 16: 66-91.

_____ and Thibaut, J. W. 1978. *Interpersonal Relations: A Theory of Interdependence*. New York: John Wiley & Sons.

Komorita, S. S., and Parks, C. D. 1995. “Interpersonal relations: Mixed-motive interaction.” *Annual Review of Psychology,* 46: 495-530.

Krueger, J., and Clement, R. H. 1994. “The truly false consensus effect: An ineradicable and egocentric bias in social perception.” *Journal of Personality and Social Psychology,* 67: 596-610.

Kuhlman, D. M., and Marhsello, A. 1975. “Individual differences in game motivation as moderators of preprogrammed strategic effects in prisoner’s dilemma.” *Journal of Personality and Social Psychology,* 32: 922-31.

Ledyard, J. 1995. “Public Goods: A Survey of Experimental Research.” In J. Kagel and A. Roth (eds.), *Handbook of Experimental Economics*. Princeton, NJ: Princeton University Press.

Levine, D. 1998. “Modeling Altruism and Spitefulness in Experiments.” *Review of Economic Dynamics* 1: 593-622.

Lichtenstein, E. 1970. “ ‘Please don’t talk to anyone about this experiment’: Disclosure of deception by debriefed subjects.” *Psychological Reports, 26,* 459-79.

Loewenstein, G., Thompson, L., and Bazerman, M. 1989. "Social utility and decision making in interpersonal contexts." *Journal of Personality and Social Psychology* 57: 426-41.

Marks, G. and N. Miller. 1987. "Ten years of research on the false-consensus effect: An empirical and theoretical review." *Psychological Bulletin* 102: 72-90.

Messick, D. and Sentis, K. 1985. "Estimating social and nonsocial utility functions from ordinal data." *European Journal of Social Psychology, 15*, 389-99.

Neale, M. and Bazerman, M. 1991. *Cognition and rationality in negotiation.* New York: Free Press.

Pillutla, M. and Murnighan, J.K. 1996. "Unfairness, Anger, and Spite: Emotional Rejections of Ultimatum Offers." *Organizational Behavior and Human Decision Processes* 68: 208-24.

Rabin, M. 1993. "Incorporating fairness into game theory and economics." *American Economic Review* 83: 1408-18.

_____. 1998. "Psychology and Economics." *Journal of Economic Literature* 36: 11-46.

Ross, L. 1977. "The intuitive psychologist and his shortcomings." In L. Berkowitz (Ed.), *Advances in experimental social psychology*, (Volume 10). San Diego, CA: Academic Press.

Ross, L., Greene, D., and House, P. 1977. "The false consensus effect: An ego-centric bias in social perception and attribution processes." *Journal of Experimental Social Psychology* 13: 279-301.

Ross, L., and Ward, A. 1995. "Psychological barriers to dispute resolution." In M. Zanna (Ed.), *Advances in experimental social psychology* (Volume 27). San Diego, CA: Academic Press.

Sally, D. 1999. "A Sympathetic Look at the Prisoners' Dilemma." Unpublished manuscript, Johnson Graduate School of Management, Cornell University.

Slonim, R. and A.E. Roth. 1998. "Learning in High Stakes Ultimatum Games: An Experiment in the Slovak Republic," *Econometrica* 66: 569-96.

Van Boven, L., D. Dunning, and G. Loewenstein. 1999. "Trading places: Egocentric empathy gaps between owners and buyers." Unpublished manuscript, Cornell University.

Weber, R. Rottenstreich, Y., Camerer, C. and Knez, M. "The Illusion of Leadership." Forthcoming in *Organizational Science.*

Figure 1. Prisoners' Dilemma game presented to participants.

		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

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

	Condition		
	Negative Personality	Control	Positive Personality
(a) OTHER'S PREF	-3.20	0.40	3.53
(b) OTHER'S MOVE	40%	87%	100%

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 = .24, sd = 4.23). It therefore ranges from –8 to +8.

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

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

Table 3. (a) Participant's own stated preference for cooperation (OWN PREF), by experimental condition. (b) Percentage of participants who choose to cooperate (OWN MOVE), by experimental condition.

	Condition		
	Negative Personality	Control	Positive Personality
(a) OWN PREF	0.60	1.87	3.80
(b) OWN MOVE	60%	73%	93%

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, sd = 3.53). It therefore ranges from –8 to +8.

Appendix

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	1. I am sincere and trustworthy. I will not lie, for whatever ends.
Me	Neutral	Not me	2. I pride myself on being highly principled. I am willing to stand by those principles no matter what the cost.
Me	Neutral	Not me	3. My sense of humor is one of my biggest assets.
Me	Neutral	Not me	4. I have above-average empathy for the views and feelings of others.
Me	Neutral	Not me	5. 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	6. I enjoy trying to persuade others to my point of view.
Me	Neutral	Not me	7. I feel if I am too honest and trustworthy, most people will take advantage of me.
Me	Neutral	Not me	7. I feel if I am too honest and trustworthy, most people will take advantage of me.
Me	Neutral	Not me	8. To persuade others, I prefer to use fear rather than trust
Me	Neutral	Not me	9. I try not to be predictable because then I can be easily manipulated
Me	Neutral	Not me	10. I love to be the aggressor. I believe I have to take the initiative if I want to accomplish my goals.
Me	Neutral	Not me	11. I believe honesty and openness are essential for maintaining good relationships.
Me	Neutral	Not me	12. In a negotiation, I believe the best outcome is one that is fair for all parties.
Me	Neutral	Not me	13. I believe one can achieve the best results in life by *cooperating* with others.
Me	Neutral	Not me	14. I believe one can achieve the best results in life by *competing* with others.
Me	Neutral	Not me	15. I believe principles are fine for some people, but sometimes they have to be sacrificed to achieve one's goals.
Me	Neutral	Not me	16. In negotiations, I try to exploit my opponent's weaknesses.
Me	Neutral	Not me	17. I believe that imposing personal discomfort is not too high a price to pay for success in negotiation
Me	Neutral	Not me	18. I believe there is nothing wrong with lying in a competitive situation, as long as I don't get caught.
Me	Neutral	Not me	19. I believe sometimes I must sacrifice the welfare of others for my own benefit.
Me	Neutral	Not me	20. I believe that the dignity and welfare of others is the most important concern for any society.