Non-Instrumental Behavior in an Environmental Public Good Game

Leif Helland*
Jon Hovi†

November 2, 2006

Abstract

This paper reports a puzzling result from an experiment based on an indefinitely repeated N-player Prisoners' Dilemma game carried out in a PC lab. The experiment used real monetary payoffs, and was conducted in the context of international cooperation to curb climate change. It was puzzling that after the experiment, a large majority of subjects reported they were at least partially motivated out of concern for the climate; however, nothing they did in the experiment could possibly have had an impact on the climate. We show that subjects acting out of concern for the climate incurred a real cost in monetary terms, and argue that many subjects' behavior in the experiment deviated quite fundamentally from instrumental rationality. Although much recent research on public goods provision questions the traditional assumption that players are purely self-regarding, the assumption of instrumental rationality is typically preserved. The results reported in this paper go some way towards challenging the validity of this assumption.

1 Introduction

Current research on public goods provision uses two types of models - self-regarding and other-regarding. Self-regarding models, which dom-

*Norwegian School of Management (BI) and CICERO. Corresponding author: Nydalsveien 37, 0442 Oslo, leif.helland@bi.no
†Department of Political Science, University of Oslo and CICERO

inated until the end of the 1980s, assume that players are motivated exclusively by self-interest. It is well known that with a finite horizon and complete information, public goods games with self-interested players typically have a unique equilibrium, in which no player contributes. With an indefinite horizon, things are more open-ended. In particular, folk theorems ensure that almost any observable behavior may be sustained as a subgame perfect equilibrium, provided that players do not discount future payoffs too severely. Obtaining empirical cutting power in the indefinite-horizon scenario thus requires more stringent - and more controversial - equilibrium refinements.

Ample experimental evidence exists on behavior in repeated public goods games. Good reviews of the literature are provided by Dawes and Thaler (1988), Ledyard (1995), and Fehr and Schmidt (1999:836-9). Almost all experiments have a finite horizon, and their lengths are typically made public knowledge before the game begins.[1] The main findings can be summarized as follows: In one-shot experiments, subjects contribute 40-60 percent of their endowment on average. With repeated play (usually 10-20 rounds) average contributions generally start out as in one-shot experiments, but drop over time (though exceptions have been documented). Contribution levels are generally low in the final period of repeated-play experiments. However, if restarted, or played by subjects having prior experience with public goods experiments, average contributions will typically start out as in one-shot experiments and drop over time. Hence, the observed patterns of behavior cannot easily be attributed to learning. Contributions tend to increase as the marginal per capita return of a contribution increases (Isaac, Walker Thomas 1984). Controlled for this factor, the effect of group size is somewhat uncertain, but appears to be positive. Pre-play communication tends to increase average contribution levels. Also, provision thresholds ("lumpy goods") increase average

[1]As far as we know, the only experiment using an indefinitely repeated public goods game is reported by Roth and Murnighan (1978). This is remarkable, considering that much of the theoretical literature on public goods provision concerns indefinitely repeated games. Roth and Murnighan use an indefinitely repeated 2x2 PD game, in which one player is an automaton. Their main finding is that reducing the continuation probability reduces average contributions. The payoff structure is kept constant in their experiment. Furthermore, players are rewarded with tickets in a final lottery for a fixed prize (10 USD). By contrast, in the experiment we report, the continuation probability is kept constant, the payoff structure is varied systematically, the game is played by groups of 5 and 10 real subjects, and payoffs are accumulated from repeated play of the stage game.

contribution levels, although this effect is uncertain. Finally, contributions tend to decrease as contribution costs increase, other things being equal.

Because several of these findings are difficult to explain with self-regarding models, from the beginning of the 1990s the use of other-regarding models has increased in the study of public goods provision. There are two types of such models. "Outcome-based" models assume that players include the welfare of other players - positively or negatively - in their own utility function (Andreoni and Miller 2000, Charness and Rabin 2002, Bolton 1991, Fehr and Schmidt 1999, Bolton and Ockenfels 2000). In contrast, "intention-based" models assume that players care not only about their opponents' payoffs, but also about their intentions.[2] In particular, intention-based models assume that players desire to reciprocate acts they perceive as kind, as well as acts they perceive as unkind (Rabin 1993, Falk and Fischbacher 2006, Dufwenberg and Kirchsteiger 1998, Charness and Rabin 2002).[3]

Both self-regarding and other-regarding models assume that players are instrumentally rational in the sense that they seek to allocate scarce resources in a way that enables them to achieve a (consistent) set of objectives, given their beliefs and other players' actions. The basic point of contention concerns the kinds of objectives players typically pursue. In this paper we report findings from a PC laboratory experiment which suggest that the subjects deviated quite fundamentally from instrumental rationality. In particular, we report a puzzling result from a Prisoners' Dilemma (PD) experiment that was conducted in the context of cooperation to curb climate change. The majority of subjects reported that their behavior in the experiment was at least partially motivated by concern for the climate. However, nothing the participants did in the experiment could possibly have had any impact on the climate. Also, subjects (fruitlessly) trying to improve the climate in the experiment suffered real monetary costs. Thus, it seems that neither self-regarding models nor other-regarding models can explain the observed behavior. In short, the subjects' behavior does not

[2]Intention-based models bring us into the realm of psychological game theory, where payoffs depend not only on outcomes of the game, but also on the players' beliefs (Geanakoplos, Pearce and Stachetti 1989).

[3]Though models based on other-regarding preferences and intentions are popular, they are far from being generally accepted. For instance, Samuelson (2005) convincingly argues that a simple model of bounded rationality can satisfactorily explain the major findings from finitely repeated experimental games.

appear to be instrumentally rational.[4]

The paper is organized as follows: Section 2 describes our experiment. Section 3 presents in detail the above-mentioned puzzling result. Finally, some ways of explaining the puzzle are discussed in section 4.

2 Experimental design

The experiment was conducted on two consecutive days, with 20 subjects participating the first day and 20 different subjects participating the second day. Subjects were recruited from students at the University of Oslo. Each subject received a show-up fee of NOK 300 (about 46 US dollar), in addition to what he or she earned in the experiment. Upon arrival each subject answered a short questionnaire on background variables (sex, age, university courses and prior knowledge of game theory). The questionnaire also included a few questions intended to tap potentially relevant dimensions of their value systems. One question was designed to tap the respondents' "green attitudes":

> Please indicate your degree of agreement or disagreement with the following statement: I am willing to sacrifice goods and services that I presently consume, if by doing so I would contribute to preserving the natural environment

The response alternatives were: "I completely disagree", "I disagree", "I find it impossible to indicate agreement/disagreement", "I agree" and "I completely agree". These alternatives were coded 0 to 4 for degree of "greenness". The subjects' responses varied from 0 to 4, with a mean of 3.05 and a standard deviation of 0.86.

Together with the questionnaires, the subjects received written instructions explaining the details of the experiment.[5] A set of control

[4]The main purpose of the experiment was to test the explanatory force of Farrell and Maskin's concept of weak renegotiation-proof equilibrium (WRPE) in an indefinitely repeated, N-player PD game. Our experimental design produces different WRPEs for different cost treatments. With a cost of 6 schillings, the WRPE has exactly 3 players contributing. With a cost of 12 schillings, it has exactly 5 players contributing. Finally, with a cost of 21 schillings, it has exactly 8 players contributing. The first two WRPEs are insensitive to variation in group size. The third is relevant only in the large group treatment. Details on the logic behind the WRPE are reported in a separate publication with other findings. In this paper we confine ourselves to discussing a puzzling by-product of our experiment.

[5]For the complete instructions see: http://home.bi.no./a0111218/INSTRUCTIONS.pdf

questions enabled us to verify that the subjects had read and understood these instructions.

The stage game implemented in our experiment was an N-player PD game. The monetary stage-game payoff for subject i was given by the following function:

$$\pi_i = (c - c_i) + ac_i + a\sum_{j \neq i}^{N} c_j,\ c_i, c_j \in \{0, c\},\ c \in \{6, 12, 21\} \qquad (1)$$

The experiment included two treatments for group size (N=5 and N=10), and three treatments for endowments (c=6 schillings, c=12 schillings and c=21 shillings). "Schillings" refers to an experimental currency with an exchange rate of 0.3 to one NOK. In each round of the experiment, the subjects could either keep or contribute their endowments. The exact value of varied with the cost treatments: $a = \frac{1}{2}$ for c=6; $a = \frac{1}{4}$ for c=12; and $a = \frac{1}{7}$ for c=21. Because $a < 1$ in all treatments, a subject motivated solely by (monetary) self-interest would have "keep" as a dominant strategy in the stage game.

The experiment consisted of a number of sessions, with a random number of rounds in each session. After each round, the computer program decided with probability 0.1 that the game would end, and with probability 0.9 that the game would continue for at least one more round. The number of rounds per session varied between 1 and 48, for an average of 12.3.

As can be seen, our experiment combines elements from an across-subjects design (group size) with elements from a within-subjects design (cost treatments). Conventionally, across-subjects designs dominate in experimental economics, although this does not seem to be well founded.[6]

A potential drawback of the across-subjects design is the risk that comparisons across subjects are blurred by uncontrolled, subject-specific differences, pertaining for instance to group heterogeneity in social

[6]Camerer (2003:42) writes: "There is a curious bias against within-subjects designs in experimental economics (not so in experimental psychology). I don't know why there is a bias, and I can't think of a compelling reason always to eschew such designs. One possible reason is that exposing subjects to multiple conditions heightens their sensitivity to the differences in conditions. This hypothesis can be tested, however, by comparing results from within- and between-subjects designs, which is rarely done." An example of a mixed design where such comparisons can be found is Sutter (2003), where the equilibrium prediction is not supported in an across-subjects design, but achieves a fair amount of support in a within-subjects design.

preferences and varying degrees of previous experience with laboratory experiments.

Since in the present article we are interested in group heterogeneity regarding social preferences, we control for within-subjects variation (through the design) as well as for the (randomly determined) composition of groups (by including intervention terms for groups).

In the experiment, all communication from us to the subjects and all interaction between the subjects took place in a computerized environment.[7] To maximize independence between sessions we used a double randomization. First, subjects were randomly distributed to groups before each session. Second, each subject was randomly assigned a subject number (an integer between 1 and N) at the start of each session. The purpose was to enable subjects to condition their decisions on other subjects' behavior in the current session (in which subjects were identifiable by their numbers but otherwise anonymous), but not on their behavior in previous sessions.

The structure of the game was made public knowledge by (i) providing subjects with payoff matrixes, information about the exchange rate of the experimental currency, and other relevant information about the game, (ii) making sure that all participants could observe that everyone received this information, and (iii) using control questions to ensure that the information was understood. Furthermore, to mirror the assumption of "almost perfect" information,[8] which underlies most theorizing about play in the indefinitely repeated Prisoners' Dilemma, a statistic was displayed on the screen of the subjects' computers at the beginning of each round. This statistic contained updated information about the history of play in all sessions up to and including the previous round, as well as own payoffs and total contributions in the group in the previous round. After entering the decision phase of a new round all subjects had continuous access to the history of play in the current session up to and including the previous round. While making decisions the subjects were reminded on screen of the cost of contributing. Each subject was also given a printout of the payoff matrix relevant to the current session.

It is well known that the context imposed on an experiment can have a significant impact on the results. However, the views on how to tackle this problem differ. One view is that the experimenter should

[7]The experiment was programmed using z-tree (Fischbacher 1999).

[8]Almost perfect information means that the history up to and including round t-1 is common knowledge in round t.

try to impose as little context as possible. The opposing view is that the idea of a context-free experiment is nave because if the experimenter does not impose a particular context, the subjects will choose their own. Therefore, "there is no 'neutral' presentation of these games, simply a variety of alternatives, so there is no way to remove the context" (Loewenstein 1999:F31). In line with the latter view, the subjects in our experiment were explicitly informed in the invitation, in the general introduction, and in their instructions, that the experiment's purpose was to test a set of hypotheses derived from a model that tries to identify conditions for international cooperation to curb climate change.

After the final session, the subjects received a second questionnaire, prompting them to indicate, on a 0-100 scale, the extent to which they had been motivated by concern for the climate, as opposed to monetary benefits, when making decisions in the experiment. For convenience we rescaled the subjects' responses, so that the scale ranged from 0 to 1 in our analysis. The responses varied from 0 to 1, with an average of 0.36 and a standard deviation of 0.29 on the rescaled weight given to the climate factor.[9]

3 A puzzling result

We present our results in three steps. First, we analyze data at the subject level (N=40). Second, we analyze data at the level of individual decisions (N=5020). Third, we include intervention terms for the (randomly formed) groups in our analysis of individual decisions.

Only five subjects (12.5 percent) reported that they placed no weight whatsoever on the climate when making decisions. Thus, for only a small fraction of the subjects can we rule out non-instrumental behavior of the type described in section 1. For these five subjects reporting a purely monetary motivation, the mean fraction of contributions - taken over all decisions made by these subjects during the experiment - was 0.34. By contrast, the corresponding mean was 0.43 for the thirty-five subjects reporting to have placed at least some weight on the climate in decision making. The small number of subjects reporting solely monetary motivation makes significance testing of differences in means unreliable. To obtain a more even distribution, we code subjects reporting a climate weighting less than 0.25 as hav-

[9]This weight will henceforth be referred to as "climate weighting".

ing "No or moderate concern for the climate", and subjects reporting a climate weighting at or above 0.25 as having a "Significant concern for the climate". When we use this convention, both subgroups become large enough to allow for statistical comparisons. In the former group the mean fraction of contributions was 0.31 (N=19), while in the latter group it was 0.51 (N=21). Table 1 shows the results of an independent samples test for the differences in means.

Table 1 here

The reported F-statistic of the Levene's test implies that the null hypothesis (that the distribution of contributions has equal variance in the two groups) may be rejected at the five percent level. Similarly, assuming heterogenous variance of contributions, the difference in the fraction of contributions between the two groups is significantly different from zero well below conventional significance levels. Thus on average, subjects reporting a significant concern for the climate contribute more often than other subjects, and this difference is statistically significant.

We also report the results of a simple multivariate analysis. Table 2 presents results of two linear regression equations (OLS estimates).

Table 2 here

In table 2, the independent variables are the climate weighting, green attitudes (using the five-point scale explained in section 2), and a dummy for group size (scoring 0 for groups with 5 subjects and 1 for groups with 10 subjects). We ran regressions on two dependent variables: (i) fraction of contributions (i.e., number of rounds in which the individual chose 'contribute' as a fraction of all rounds in which he or she was a player),[10] and (ii) average monetary payoff in the rounds played.

The results can be summarized as follows: Climate weighting significantly affects the fraction of contributions. Other things being equal, increasing the climate weighting from zero to one increases the average fraction of contributions from zero to 0.62, a sizeable increase. Group size and green attitudes do not significantly contribute to ex-

[10]This means that if a subject were to play 60 rounds, and play 'contribute' in 15 rounds, then that subject would score 0.25 on this dependent variable, which is henceforth referred to as "fraction of contributions".

plaining the variation in contributions.

The climate weighting also significantly explains average monetary payoffs. Other things being equal, the average monetary payoff per round is 4.31 schillings lower for subjects having a climate weighting of one than for subjects having a climate weighting of zero, again a sizable effect. Thus, acting out of concern for the climate was costly. Unsurprisingly, group size now matters because a given fraction of contributions will produce higher monetary payoffs in a large group than in a small one. Green attitudes do not contribute significantly to explaining the variance in average monetary payoffs.

Interestingly, the correlation between climate weighting and green attitudes is moderate - Pearson's r =0.26 (and only just significant at the ten-percent level). Table 3 shows the bivariate distribution for green attitudes and climate weighting (now measured as a trichotomy).

Table 3 here

From table 3 we see that 33 of the 40 subjects (83 percent) report strong green attitudes. More than half of these "green" subjects report a fairly low climate weighting - 0.33 or less. This, of course, is by no means an inconsistency. A subject may well have had strong green attitudes, and yet realized that nothing he or she did in the experiment could have helped the environment in any way. Of the subjects with medium to high climate weighting, all except one have strong or very strong green attitudes. It is conceivable that some subjects in this group may have let their strong green attitudes guide their behavior in the experiment; we discuss this possibility further in section 4. Given the distribution in table 3, it is not surprising that green attitudes fail to explain variation in contributions and payoffs. Strong green attitudes are common - not only among subjects reporting a significant concern for the climate, but also among other subjects.

We now turn to the results of an analysis at the level of individual decisions. Table 4 shows the results of a logistical regression used to estimate the probability of a contribution, conditioned on eight independent variables. The independent variables include climate weighting, green attitudes, and the dummy for group size, as defined above. In addition, there are two dummies for cost. The first scores one if the cost is 12 schillings ("high cost") and zero otherwise. The second scores one if the cost is 21 schillings ("very high cost") and zero otherwise. Thus, a cost of 6 schillings ("low cost") serves as the baseline

category and scores zero on both "high cost" and on "very high cost". The dummy for gender is coded one for male and zero for female. The final two variables are session number (per day) and round number (per session), which are included in order to control for dynamics. Using the same set of independent variables we also ran a linear regression with monetary payoffs per round as dependent.

Table 4 here

The results from the analysis of individual decisions corroborate the findings in table 2. Climate weighting significantly explains the probability of contributing, controlled for cost structure, group size, gender, green attitudes and dynamics. Also, acting out of concern for the climate is clearly costly. The effect of green attitudes is now statistically significant at conventional levels. This holds regardless of whether probability of contributing or monetary payoffs per round is used as the dependent variable. However, in both cases the controlled effect of green attitudes is miniscule compared to that of climate weighting, and with more than 5000 observations, one should not put too much emphasis on statistical significance alone. High costs and small groups imply lower contribution frequencies. Very high costs increase monetary payoffs per round, and so do large groups. The dynamic effects (given by the coefficients of rounds and session numbers) are small, but subjects contribute slightly more (rather than less) in later sessions.

Figure 1 displays a pair of probability curves drawn on the basis of the logistical regression estimates in table 3. The curves are drawn for a male subject in a big group (10 subjects), in a high cost (12 schillings) treatment, and making a decision in round 15 of session 6. The two curves represent subjects having strong (score 4) and weak (score 0) green attitudes, respectively, and show how the probability that these subjects contribute varies with the climate weighting. Whereas the difference in the probability of contributing due to differences in green attitudes is never above 7 percentage points, the difference in the probability of contributing due to differences in climate weighting is near 70 percentage points. The fact that the curves are fairly parallel suggests that there is no significant interaction effect between green attitudes and climate weighting.

Figure 1 here

Our findings are summarized in the following observations:

Observation 1 Only a small fraction of the subjects reported no concern whatsoever for the climate (a climate weighting of zero) when making decisions in the experiment.

Observation 2 The variance in climate weighting cannot be explained by variance in green attitudes.

Observation 3 On average, subjects with a high climate weighting contributed more often than subjects with a low climate weighting.

Observation 4 Subjects acting out of concern for the climate paid a price in terms of reduced monetary payoffs.

4 Discussion

Given that nothing the subjects did in the experiment could possibly influence the climate, it is puzzling that a large majority of subjects nevertheless reported that their behavior was motivated by climate-related considerations. We now discuss some possible explanations for this apparent puzzle. Section 4.1 considers other-regarding preferences in some detail, whereas section 4.2 looks more briefly at a number of other (mostly methodological) possibilities.

4.1 Other-regarding preferences

Can the subjects' (fruitless) attempts to improve the climate be accounted for by models based on other-regarding preferences? Recall that there are two types of such models. Intentions-based models are problematic in the N-player PD game, where it is impossible for a player to reciprocate a particular player's unfriendly (or friendly) behavior without also punishing (or rewarding) other players. Intentions-based models have so far been worked out only for games in which punishments and rewards can be accurately targeted. For this reason we focus on outcome-based models. The most elaborate models of this type are developed by Fehr and Schmidt (1999) and Bolton and Ockenfels (2000). For the parameters in our experiment, Bolton and Ockenfels's model rules out stage-game equilibria where only a

fraction of the subjects contribute to the public good.[11] Therefore we concentrate on the model proposed by Fehr and Schmidt (FS).

In the FS model player j's utility (i) increases as player j's own payoff increases, (ii) increases as an opponent's payoff increases if that opponent's payoff does not exceed that of player j, and (iii) decreases as an opponent's payoff increases if that opponent's payoff exceeds that of player j. This utility function might rationalize positive as well as negative actions towards an opponent. The model allows for heterogeneity of preferences in the pool of subjects, and other-regarding subjects might condition their behavior on the distribution of preferences in the player group. This makes preference heterogeneity a critical factor for the number of contributions in equilibrium.

The FS model identifies two stage-game equilibria in the N-player PD game. In the first equilibrium, no player contributes. In the second, there are positive contribution levels. While the no-contribution equilibrium always exists, the positive-contribution equilibrium exists only for certain (unlikely) parameter values. Loosely, it exists if a sufficiently large fraction of players do not have defection as a dominant strategy in the stage game. Whether a player has defection as a dominant strategy depends on (i) the marginal productivity of a contribution to the public good, and (ii) the utility the player derives from helping opponents that are less well off. The exact conditions for the existence of an equilibrium with positive contributions in the FS model are described in the appendix.[12]

An interesting question is: Can other-regarding players achieve higher contribution levels than self-regarding players in the indefinitely repeated PD game, and if so, under what circumstances? It is well known that full cooperation among self-regarding players can be sustained as a subgame perfect equilibrium in the indefinitely repeated game, provided that the discount factor is high enough.

Proposition 1. *Self-regarding players will almost always be able to achieve contribution levels as high as those of other-regarding players in the indefinite game.*

Proof. See appendix. □

[11] More precisely, for public goods games with $N \geq 2$ and $a < 1$ no positive contribution equilibrium exists in the Bolton and Ockenfels model.

[12] The more general conditions of Fehr and Schmidt's (1999) proposition 5 address a larger class of public goods games than the PD game does.

If (as suggested by proposition 1) the behavior of self-regarding and other-regarding subjects is practically indistinguishable in our experiment, our puzzling result cannot be explained by the presence of other-regarding preferences. To check whether the effect of climate weighting on contributions disappears once we allow for other-regarding preferences, we control for variance due to group heterogeneity (the main determinant of contributions in the FS model). Table 5 displays regressions with dummies for groups included. To allow for possible learning effects and for convergence of behavior within groups, we ran these regressions consecutively on sessions with more than 20 rounds of play, more than 10 rounds of play, more than 5 rounds of play, and for all sessions. When we are using dummies for groups, including costs in the regressions (unsurprisingly) leads to severe problems with multicolinearity.[13] Costs were therefore left out. Otherwise, the independent variables are as in table 4.

Table 5 here

Note first that the effect of climate weighting in the regression for all sessions is almost unchanged from the effect obtained in table 4. The explanatory force of climate weighting is basically unchanged after controlling for heterogeneity over groups. Thus, the effect of climate weighting is not spurious. Nor is it merely mediating the effect of social preferences. Second, the more time we allow for convergence, the stronger the effect of climate weighting. This finding suggests that subjects' non-instrumental behavior becomes more (not less) pronounced with experience. Finally, controlling for heterogeneity over groups, the green-attitudes variable has no effect on the probability of making a contribution.

The positive effect of playing in later sessions is not particularly robust. There is, however, a modest tendency that the probability of making a contribution is greater in late rounds than in early ones. Finally, the gender dummy is fairly stable over the regressions. The signs are the same and the coefficients have about the same magnitudes as

[13] Qualitatively similar results are obtained by running a random effects logistical regression model on contributions using the same independent variables as in table 4, using subject ID as panel variable and total number of rounds as time variable. The random effects formulation allows for the cost dummies to enter. As in the table 4 regression, these dummies give rise to significantly negative coefficients. Results are provided upon request.

in table 4. This leads to our final two observations:

Observation 5 The effect of climate weighting is not significantly reduced when controlling for group heterogeneity.

Observation 6 If anything, the effect of climate weighting increases as the subjects gain experience playing within a particular group.

4.2 Other explanations

Based on the above discussion we conclude that our puzzling result does not seem to be explained by other-regarding preferences. Therefore, we briefly consider other explanations:

Ignorance: One possibility is that many subjects simply did not realize that their behavior in the experiment had no impact on the climate. While we have no hard evidence that enables us to completely rule out this possibility, we find it extremely unlikely that subjects believed that their actions in a simple experiment carried out in a PC laboratory could have an impact on the climate.

Leading question: Pre-play instructions made it clear that the experiment's purpose was to test a model designed to explain international cooperation to curb climate change. Hence, the context explicitly imposed on the experiment was one of cooperation on climate change. Therefore, subjects may have felt they were expected to report a positive climate weighting. In other words, the question posed to the subjects after the experiment may have been leading. There is certainly something to this. However, if subjects who reported a positive climate weighting did this simply because the question was leading, one would not expect a strong correlation between climate weighting and the subjects' propensity to play "contribute" during the experiment. Thus, the leading-question explanation is not very plausible either.

Modest monetary incentives: The maximum aggregate payoff that could be achieved by a subject in the experiment was around 150 US dollar. Many students are likely to consider this sum significant, especially because the experiment lasted only around two hours. Nevertheless, the monetary incentives in each round were undeniably modest. Thus, it is possible that some subjects largely disregarded the monetary incentives and acted instead out of other motivations. However, one might question -both on theoretical grounds and on the basis of evidence from experiments with explicit high stake treatments - whether

there actually is an effect of modest stakes on behavior in experiments (c.f. Camerer 2000:60-62 for a discussion).

Boredom: The experiment involved the subjects in repeated decisions that might quickly become routine or even boring. We cannot exclude the possibility that this may have had a bearing on their behavior. Again, however, it is difficult to see why boredom would produce a strong link between climate weighting and the propensity to contribute. Also, if such a link existed, it should have become more and more apparent as the experiment progressed. However, no such dynamic pattern is discernible in the data.

Expressive motivation: March and Olsen's (1989) useful distinction between the logic of consequentiality and the logic of appropriateness is well known. As we have seen, it is difficult to account for our puzzling result in terms of the logic of consequentiality. The observed behavior is not consistent with models based on either self-regarding or other-regarding preferences. Nor can the behavior be explained in terms of concern for climate-related consequences. However, it is also difficult to provide a convincing answer in terms of the logic of appropriateness. For the latter type of logic to work, behavior must be observable by other players. We went far in our design of the experiment to make sure that behavior would be anonymous. There was no way that a subject could be identified by others as a person that behaved in a socially or morally appropriate way.

Ex-post rationalization: While the statistical analysis in the previous section sees subjects' experimental behavior as being determined (at least partly) by climate weighting, it is conceivable that it was in fact the other way around, i.e. that climate weighting was determined by the subjects' behavior. Recall that the subjects were asked about climate weighting only after the experiment had been completed. Although at that stage each subject had not yet been informed about his or her aggregate monetary payoff, most subjects probably had a reasonably good idea about their own performance. Thus, subjects that did well (in monetary terms) might simply have tried to highlight their own "success" by stressing that making money was their main objective in the experiment. Conversely, subjects that made less money might have used the opportunity to excuse their "failure" by emphasizing that they had not tried to make money in the first place. Taking the opportunity to make this "excuse" may have seemed all the more attractive (to those that did relatively poorly in monetary terms) because being motivated out of concern for the climate stood out as

the morally superior alternative. Hence, reporting concern for the climate rather than a desire to maximize monetary gains might have been a way for subjects that did poorly (in monetary terms) to justify their performance afterwards in a way that seemed morally attractive as well as logically consistent with their own behavior. However, plausible as this explanation might seem, it is difficult to reconcile with instrumental rationality. The simple reason is that, as we have stated repeatedly, nothing the subjects did in the experiment could possibly have had an impact on the climate. It is thus hard to see how it could be instrumentally rational to incur real monetary costs if one were aware that one's efforts to improve the climate in this experiment were completely fruitless.

5 Conclusion

In this paper we report a puzzling result from an experiment based on an indefinitely repeated N-player PD game, carried out in a PC lab. The experiment used real monetary payoffs, and was conducted in the context of international cooperation to curb climate change. It is puzzling that after the experiment, a large majority of subjects reported they were motivated out of concern for the climate, although nothing they did in the experiment could have had an impact on the climate. Thus, the subjects' behavior in the experiment seems to deviate quite fundamentally from instrumental rationality. Having described this result in some detail, we discussed a number of possible explanations. Our conclusion is that none of these explains the puzzling result. Although much recent research on public goods provision questions the traditional assumption that players are purely self-regarding, the assumption of instrumental rationality is typically preserved. The results reported in this paper go some way towards challenging the validity of this assumption.

A Proof of proposition 1

Denote the N-player PD stage game by G. Assume $0 < a < 1$, $c_i = \{0, c\}$, $c > 0$, and let $0 \leq k \leq (n-1)$ be the number of defecting players exempting player i. Let α_i be a measure of the utility loss from disadvantageous inequality, while β_i is a measure of the utility loss from advantageous inequality. Assume that $\alpha_i \geq \beta_i$ and

$0 \leq \beta_i < 1$. If player i contributes in G no player will be behind him, and there is no utility loss from advantageous inequality. His utility according to the FS model becomes:

$$U_i(c_i - c) = ac + a(n-1-k)c - \frac{\alpha_i}{n-1}kc \tag{2}$$

If player i defects in G no player will be ahead of him, and there is no utility loss from disadvantageous inequality. His utility according to the FS model becomes:

$$U_i(c_i - c) = c + a(n-1-k)c - \frac{\beta_i}{n-1}(n-k-1)c \tag{3}$$

FS prove the following (1999: Proposition 4): If $a + \beta_i < 1$ player i's dominant strategy in G is to defect. Let $0 \leq k^\circ \leq N$ denote the number of players with defection as a dominant strategy in G. Universal defection is then the unique equilibrium in G if $\frac{k^\circ}{n-1} > \frac{a}{2}$.

If $\frac{k^\circ}{n-1} < \frac{a+\beta_i-1}{a+\beta_i}$ an equilibrium with positive contributions also exists in G.[14] In this equilibrium k° players defect, while $(n - k^\circ)$ players contribute. We focus on the universal defection equilibrium in G.[15]

Define υ_i as the periodic payoff on the equilibrium path in the δ-discounted indefinitely repeated PD game. Define $\lambda_i \equiv \max_{s_i} u_i(s'_{-i})$ as the payoff associated with the most profitable single-period deviation from the equilibrium path. Let e_i denote the payoff in the G equilibrium. The folk theorem states that a strategy vector can be constructed such that any $\upsilon_i \geq e_i$ is implemented as a subgame perfect equilibrium (SPE) in the repeated game, if for every player $\delta > \frac{\lambda_i - \upsilon_i}{\lambda_i - e_i}$.[16] The theorem assumes that a deviation is punished by reversion to the G equilibrium in all subsequent periods. This gives every player a per-period payoff of e_i.

In what follows we focus on sustaining universal contribution as a SPE in the repeated game. In other words, we consider the case

[14] By comparison, Bolton and Ockenfels's model rules out stage-game equilibria where a fraction of the players contribute if $a < 1$ and $N \geq 2$.

[15] We need not check the SPE having reversion to a positive-contribution equilibrium in G as a threat in order to highlight proposition 1, since when this equilibrium exists then so does the universal-defection equilibrium.

[16] Friedman (1971).

where $v_i = anc$. Like Friedman, we assume that after a defection all players revert to the universal-defection equilibrium in G. For every player $e_i = c$. It may be noted that players with $a + \beta_i > 1$ have no incentive to deviate from the equilibrium path, even in the short term. Hence, we concentrate on players for whom $a + \beta_i < 1$. For these players $\lambda_i = c + a(n-1)c - \beta_i c$ and the conditions for Nash equilibrium are that, for all i $(i = 1, 2, ...n)$,

$$\delta \geq \frac{1 - a - \beta_i}{a(n-1) - \beta_i} \tag{4}$$

The right-hand side of condition (4) is decreasing in β_i, the only parameter that varies across players. The effective condition for Nash equilibrium is therefore that (4) is satisfied for the player with the lowest β_i.

For this equilibrium to be subgame perfect, we must also require that the threatened punishment for defection is credible in the sense that it is individually rational to revert to the G equilibrium, given that all other players do so. That this is rational follows immediately from the fact that we are dealing with a Nash equilibrium in G (i.e., in the stage game). Satisfying (4) for the player with the lowest β_i (the player that is least concerned about being ahead of other players) is therefore a necessary and sufficient condition for a SPE.

Consider a homogenous group of purely self-regarding players. The utility function (assuming linearity in monetary payoff) for these players is given by (1), meaning that $\lambda_i = c + a(n-1)c$, and $e_i = c$. The Nash condition then becomes:

$$\delta \geq \frac{1 - a}{a(n-1)} \tag{5}$$

Again it is trivial that the threat of reverting to the Nash equilibrium in G is individually rational, so that (5) is a necessary and sufficient condition for a SPE.

FS (1999:Table 3) provi de simple discrete distributions for α_i and β_i, based on numerous experiments with the ultimatum game.[17] Comparing (4) and (5), and using the distributions in FS, makes the statement in proposition 1 precise. FS estimate that $\beta_i = 0$ for 0.30 of the

[17]FS use these estimates to calculate the probability of there being a positive-contribution equilibrium in the public good game.

individuals in the subject pools used. Thus, in our experiment the expected probability of there being at least one subject with $\beta_i = 0$ is (by the binomial distribution) 0.83 in the small group treatment, and 0.97 in the large group treatment. The probability of there being at least one subject with $\beta_i = 0$ in our groups is in other words very high. With at least one subject for whom $\beta_i = 0$, (5) becomes the relevant condition for universal cooperation to be a SPE. This is the quantification of "almost always" in proposition 1.

REFERENCES

Andreoni, J. and J. Miller (2000): Giving according to GARP: An Experimental Test of the Consistency of Preferences for Altruism. *Econometrica* 70(2): 737-53.

Bolton, G. (1991): A Comparative model of bargaining: Theory and Evidence. *American Economic Review* 81: 1096-1136.

Bolton, G. and A. Ockenfels (2000): ERC: A theory of equity, reciprocity, and competition. *American Economic Review* 90: 166-93.

Camerer, C. (2000): *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton: Princeton University Press.

Charness, G. and M. Rabin (2002): Understanding social preferences with simple tests. *Quarterly Journal of Economics* 117:817-69.

Dawes, R. and R. Thaler (1988): Cooperation. Journal of Economic Perspectives 2(3): 187-97.

Dufwenberg, M. and G. Kirchsteiger (1998): *A Theory of Sequential Reciprocity*. Tilburg Center for Economic Research. Discussion paper 9837.

Falk, A. and U. Fischbacher (2006): A Theory of Reciprocity. *Games and Economic Behavior* (forthcoming).

Fehr, E. and K. Schmidt (1999): A theory of fairness, competition and cooperation. *Quarterly Journal of Economics* 114: 817-68.

Fischbacher, U. (1999): *z-Tree - Zurich Toolbox for Readymade Economic Experiments - Experimenter's Manual*. Working Paper Nr. 21. Institute for Empirical Research in Economics, University of Zurich.

Friedman, J. (1971): A noncooperative equilibrium for supergames. *Review of Economic Studies* 38: 1-12.

Geanakoplos J., D. Pearce and E. Stachetti (1989): Psychological Games and Sequential Rationality. *Games and Economic Behavior* 1: 60-79.

Isaac, R.J.S., J. Walker and S. Thomas (1984): Divergent Evidence on Free Riding: An Experimental Examination of Possible Explanations. *Public Choice* 43:113-49.

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

Loewenstein, G. (1999): Experimental Economics from the Vantage-Point of Behavioral Economics. *The Economic Journal* 109 (February): F25-F34.

March, J. and J. OLSEN (1989): *Rediscovering Institutions: The Organizational Basis of Politics*. New York: The Free Press.

Rabin, M. (1993): Incorporating Fairness into Game Theory and Economics. *American Economic Review* 83:1281-1302.

Roth, A. and J. Murnighan (1978): Equilibrium Behavior and Repeated Play of the Prisoners Dilemma. *Journal of Mathematical Psychology* 17: 189-98.

Samuelson, L. (2005): Foundations of Human Sociality: A review essay. *Journal of Economic Literature* 43(2): 488-97.

Sutter, M. (2003): The Political Economy of Fiscal Policy: An Experimental Study on the Strategic Use of Deficits. *Public Choice* 116: 313-32.

Table 1: Independent sample test for difference in means between subjects having no or moderate concern for the climate, and subjects having significant concern for the climate.

	Difference in means	P-value for difference in means	F-statistics homogenous variance (p-value)
Homogenous Variance	.20	.007	4.28 (.046)
Heterogenous Variance	.20	.006	_

Table 2: Regression equations. Subjects. Coefficients. (P-values).

Independents	**Dependent: Fraction of contributions**	**Dependent: Average monetary payoffs**
Constant	.15 (.133)	11.70 (.000)
Climate weighting	.62 (.000)	-4.31 (.000)
Group size	-.03 (.563)	7.93 (.000)
Green attitude	.01 (.318)	.29 (.783)
F-Statistics	14.8 (.000)	63.0 (.000)
Adjusted R^2	.52	.83
N	40	40

Table 3: Distribution of subjects on climate weighting and green attitudes

	Climate weighting		
Green attitudes	**Low (0, .33)**	**Medium (.33, .66)**	**High (.67, 1.00)**
Very weak (0)	1	0	0
Weak (1)	1	0	0
Undecided (2)	4	0	1
Strong (3)	12	6	3
Very strong (4)	5	3	4

Table 4: Regression equations. Individual decisions. Coefficients. (P-values).

	Logistical regression: Contribution	OLS: Monetary payoffs per round
Constant	-2.26 (.000)	12.38 (.000)
Climate weighting	3.47 (.000)	-5.16 (.000)
Dummy high cost	-1.58 (.000)	.08 (.657)
Dummy very high cost	-1.78 (.000)	4.80 (.000)
Dummy group size	.41 (.000)	6.78 (.000)
Dummy gender	.62 (.000)	-1.54 (.000)
Green attitudes	.08 (.037)	.21 (.018)
Round number	.001 (.859)	-.04 (.000)
Session number	.04 (.001)	.15 (.000)
χ^2	1256 (.000)	-
Percent correct	74.6	-
F-statistics	-	520 (.000)
Adjusted R^2	-	.45
N	5020	5020

Table 5: Regression equations controlled for group dummies. Individual decisions. Coefficients. (P-values).

	Logistical regression: Contributions			
	Sessions with rounds ≥ 20	**Sessions with rounds ≥ 10**	**Sessions with rounds ≥ 5**	**All sessions**
Constant	5.12 (.016)	-.01 (.998)	-.85 (.641)	-1.17 (.787)
Climate weighting	4.21 (.000)	3.71 (.000)	3.61 (.000)	3.55 (.000)
Round	.08 (.000)	.04 (.000)	.02 (.000)	0.00 (.000)
Session	-.95 (.000)	-.65 (.208)	-.42 (.298)	-.15 (.919)
Gender	.67 (.000)	.77 (.000)	.82 (.000)	.81 (.000)
Green attitude	-.02 (.863)	-.03 (.637)	-.03 (.546)	-.05 (.259)
# Group Dummies	8	30	40	66
Percent correct	80.2	77.7	78.2	76.7
χ^2	481 (.000)	867 (.000)	1335 (.000)	1791 (.000)
N	1060	2320	3580	5020

Figure 1: Probability of contributing as a function of the climate weighting, shown for strongest (4) and weakest (0) green attitudes.