Università Ca' Foscari Venezia

Dottorato di ricerca in Economia, 22° ciclo
(A. A. 2006/2007 – A.A. 2008/2009)

Essays on Fairness Heuristics and Environmental Dilemmas

SETTORE SCIENTIFICO-DISCIPLINARE DI AFFERENZA: SECS- P/06

Tesi di dottorato di Alessandro Tavoni , 955272

Coordinatore del dottorato
Prof. Agar Brugiavini

Tutore del dottorando
Prof. Carlo Carraro

Essays on Fairness Heuristics and Environmental Dilemmas*

Alessandro Tavoni

Advanced School of Economics, University of Venice
Cannaregio 873, 30121 Venice, Italy

June 2010

Abstract

The issues explored in this work concern individual behaviour and its departure from the rationality paradigm. While different in terms of underlying methodology, the chapters share the unifying theme of fairness as a guiding principle for human behaviour, as well as a focus on its relevance for environmental dilemmas.

*At the academic and personal level, I owe much to my supervisor Carlo Carraro. Along the path, many others have significantly contributed: among those based in Venice, Michele Bernasconi and Massimo Warglien have been supportive and insightful, not to mention my colleagues. Abroad, I have felt as if at home thanks to Princeton's Simon Levin and Maja Schlüter, Autonoma of Barcelona's Jeroen van den Bergh and Giorgos Kallis, and ZEW's Andreas Löschel and Astrid Dannenberg. Regardless of where I was, my parents, Giusi, my brother, Vale and Tupo have always been with me.

Introduction to the dissertation

The last decades have witnessed what some have termed the complexity revolution in economics (Rosser et al., 2010); the authors claim that "modern economics can no longer usefully be described as 'neoclassical', but is much better described as complexity economics", which "embraces rather than assumes away the complexities of social interaction." Addressing complexity is a task that requires multiple angles of attack; combining theoretical investigations and applied techniques appears to be a promising approach to inform the scientific and policy debates. In particular, evolutionary game theory has been credited for redefining how institutions are integrated into the analysis, behavioural economics for redefining how rationality is treated and experimental economics for changing the way economists think about empirical work (Colander et al., 2004). The present thesis sets to draw from these branches of economics in an effort to analyze individual behaviour with the lens which is most appropriate given the research question at hand. The essay presented in the first chapter tackles the issue of rationality by confronting the predictions of several game theoretic equilibrium concepts with empirical observations concerning 2x2 games; particular attention is given to the role of inequality aversion in explaining human behaviour. The following chapter utilizes evolutionary game theory, and the assumption that strategies of successful agents spread in the population and replace strategies of the less successful, to provide insights on the behaviour of harvesters of a common pool resource facing costly ostracism insofar they extract an amount which exceeds the socially acceptable level. The third essay relies on the experimental method to gain insights on the drivers of cooperation among subjects faced with a social dilemma framed in terms of dangerous climate change: whether to contribute to the public good, reducing their savings but increasing the likelihood of successful group coordination, or to act selfishly in the hope that others will compensate. The last section adds a brief reflection on the insights from the analyses presented throughout the thesis, as well as some concluding remarks and ideas for related work.

When fairness bends rationality: incorporating inequity aversion in models of regretful and noisy behavior[1]

Abstract

Substantial evidence has accumulated in recent empirical works on the limited ability of the Nash equilibrium to rationalize observed behavior in many classes of games played by experimental subjects. This realization has led to several attempts aimed at finding tractable equilibrium concepts which perform better empirically.Two such examples are the impulse balance equilibrium (IBE - Selten and Chmura, 2008), which introduces a psychological reference point to which players compare the available payoff allocations, and a model of stochastic choice such as the quantal response equilibrium (QRE - McKelvey and Palfrey, 1995). This paper is concerned with advancing and confronting with empirical data two concepts: equity-driven impulse balance equilibrium (EIBE) and equity-driven quantal response equilibrium (EQRE): both introduce a distributive reference point to the corresponding established stationary concepts. The explanatory power of the considered models leads to the following ranking, starting with the most successful in terms of fit to the experimental data: EQRE, IBE, EIBE, QRE and Nash equilibrium.

Keywords: Fairness, Inequity aversion, Aspiration level, Impulse balance, Quantal Response, Behavioral economics, Experimental economics

[1]This chapter has benefited from the input of Massimo Warglien and the dataset kindly made available to everyone by Reinhard Selten and Thorsten Chmura.

1 Introduction

In recent years experimental economists and psychologists have accumulated considerable evidence that steadily contradicts the self-interest hypothesis embedded in equilibrium concepts traditionally studied in game theory, such as Nash's. The lab and field evidence, together with theoretical contributions from students of human behavior belonging to fields as diverse as biology and sociology, suggests that restricting the focus of analysis to the strategic interactions among perfectly rational players (exhibiting equilibrium behavior) can be limiting, and that considerations about fairness and reciprocity should be accounted for.[2]

In fact, while models based on the assumption that people are exclusively motivated by their material self-interest perform well for competitive markets with standardized goods, misleading predictions arise when applied to non-competitive environments, for example those characterized by a small number of players (cf. Fehr and Schmidt, 1999) or other frictions. For example, Kahneman et al. (1986) find empirical results indicating that customers are extremely sensitive to the fairness of firms' short-run pricing decisions, which might explain the fact that some firms do not fully exploit their monopoly power. One prolific strand of literature on equity issues focuses on relative measures, in the sense that subjects are concerned not only with the absolute amount of money they receive but also about their relative standing compared to others. Bolton (1991) formalized the relative income hypothesis in the context of an experimental bargaining game between two players. Kirchsteiger (1994) followed a similar approach by postulating envious behavior. Both specify the utility function in such a way that agent i suffers if she gets less than player j, but she is indifferent with respect to j's payoff if she is better off herself. The downside of the latter specifications is that, while consistent with the behavior in bargaining games, they fall short of explaining observed behavior such as voluntary contributions in public good games.[3]

[2]For supporting arguments see, among the many available literature reviews, the updated one provided in Gowdy (2008).

[3]A substantial departure from the models considered here, which are solely based on

A more general approach has been followed by Fehr and Schmidt (1999), who instead of assuming that utility is either monotonically increasing or decreasing in the well being of the other players, model fairness as self-centered inequality aversion. Based on this interpretation, subjects resist inequitable outcomes, that is they are willing to give up some payoff in order to move in the direction of more equitable outcomes. More specifically, a player is altruistic towards other players if their material payoffs are below an equitable benchmark, but feels envy when the material payoffs of the other players exceed this level. To capture this idea, the authors consider a utility function which is linear in both inequality aversion and in the payoffs. Formally, for the two-player case ($i \neq j$):

$$U_{ij} = x_i - \alpha_i \, max \{x_j - x_i, 0\} - \beta_i \, max \{x_i - x_j, 0\} \tag{1}$$

In (1), U_{ij} is the subjective utility of player i when matched with player j, x_i, x_j are player i and player j's payoffs, respectively, and β_i, α_i are player i's inequality parameters satisfying the following conditions: $\beta_i \leq \alpha_i$ and $0 \leq \beta_i \leq 1$. The second term in the right-hand side of equation (1) is the utility loss from disadvantageous inequality, while the third term is the utility loss from advantageous inequality. Due to the above restrictions imposed on the parameters, for a given payoff x_i, player i's utility function is maximized at $x_i = x_j$, and the utility loss from disadvantageous inequality ($x_i < x_j$) is larger than the utility loss incurred if player i is better off than player j ($x_i > x_j$). Notice that the asymmetric behavior implied by the constraint $\beta_i \leq \alpha_i$ as well as the assumption that an individual may not experience spite towards a worse-off opponent ($\beta_i \geq 0$) or may not be willing to throw away money so as to reduce disparities ($\beta_i \leq 1$), may

subjective considerations to differences in payoffs, is represented by models where agents' responses are also driven by the motivations behind the actions of the other player. This is the case for Falk and Fischbacher (2006), as well as Levine (1998). While without doubt one can argue that our social interactions are to some extent influenced by judgments we hold on others, these efforts inevitably run into the questionable assumption of perfect (or high degree of) knowledge of the preferences. For this reason, we restrict attention here to more parsimonious models that nevertheless account for reference dependence in several dimensions, as will be explained below.

not be justified in all domains, as will be discussed in greater detail in the concluding section. The choice of retaining the above restrictions has been taken on the grounds of facilitating comparisons with the standard model, as well as in order to impose structure on the parameters and avoid to advance concepts whose predictive performance is motivated merely by the inclusion of free parameters.

Fehr and Schmidt (1999) show that the interaction of the distribution of types with the strategic environment explains why in some situations very unequal outcomes are obtained while in other situations very egalitarian outcomes prevail. In fact, the utility function in (1) has proved successful in many applications, mainly in combination with the Nash equilibrium, and will therefore be employed in this study, although in conjunction with different equilibrium concepts. In referring to the social aspects introduced by this utility function, one could think of inequality aversion in terms of an interactive framing effect (reference point dependence): this is one way to depart from considerations of sole efficiency and move towards a concept that embodies distributive concerns on the players' part.[4]

Recognizing the importance of psychological introspection on own achievement, distributive concerns with relative payoffs as well as cognitive limitations in steering individuals' behavior,we propose two equilibrium models with the aim of accounting for multiple facets determining individual behavior, such as fairness motives, regret considerations and unobserved factors. The first two are tackled with what we term equity-driven impulse balance equilibrium, while fairness considerations and noisy behavior are the main ingredients of the other model. In the next section, the main features of the impulse balance equilibrium will be introduced, while the remainder of the paper is concerned with advancing two equity-driven concepts: Section 3 deals with the proposed modification of IBE and its ability to match observed behavior by individuals playing experimental games, while Section 4 is concerned with equity-driven quantal response equilibrium and its fit to the experimental data. Section 5 provides a discussion of the results.

[4]See Kahneman and Tversky (1979) for the pioneering work that introduced the standard reference dependence concept.

2 The "psychological" reference point

The predictive weaknesses of the Nash equilibrium are pointed out, among others, by Erev and Roth (1998), who study the robustness and predictive power of learning models in experiments involving at least 100 periods of games with a unique equilibrium in mixed strategies. They conclude that the Nash equilibrium prediction is, in many contexts, a poor predictor of behavior, while claiming that a simple learning model can be used to explain, as well as predict, observed behavior on a broad range of games, without fitting parameters to each game. A similar approach, based on within-sample and out-of-sample comparisons of the mean square deviations, will also be employed in this paper to assess to what extent is the proposed model able to fit and predict the frequencies of play recorded by subjects of an experiment involving several games with widely varying equilibrium predictions.

Based on the observation of the shortcomings of mixed Nash equilibrium in confronting observed behavior in many classes of games played by experimental subjects, an alternative tractable equilibrium has been suggested by Selten and Chmura (2008). Impulse balance equilibrium is based on learning direction theory (Selten and Buchta, 1994), which is applicable to the repeated choice of the same parameter in learning situations where the decision maker receives feedback not only about the payoff for the choice taken, but also for the payoffs connected to alternative actions. If a higher parameter would have brought a higher payoff, the player receives an upward impulse, while if a lower parameter would have yielded a higher payoff, a downward impulse is received. The decision maker is assumed to have a tendency to move in the direction of the impulse. IBE, a stationary concept which is based on transformed payoff matrices as explained below, applies this mechanism to 2x2 games. The probability of choosing one of two strategies (for example to move Up) in the considered games is treated as the parameter, which can be adjusted upward or downward.[5] It is assumed that the sec-

[5]Section 3 and the Appendix provide more detail on the experimental setup utilized here.

ond lowest payoff in the matrix is an aspiration level determining what is perceived as profit or loss (with losses weighing twice as much as gains). In impulse balance equilibrium expected upward and downward impulses are equal for each of both players simultaneously.

The main result of the paper by Selten and Chmura (2008) is that, for the games they consider, impulse balance theory has a greater predictive success than the other stationary concepts they compare it to: Nash equilibrium, action-sampling equilibrium, payoff-sampling equilibrium and quantal response equilibrium. While having the desirable feature of being a parsimonious parameter-free concept as the Nash equilibrium, and of outperforming the latter, the aspiration level framework (to be described) has the less appealing featuring of requiring the use of transformed payoffs in place of the original ones for the computation of the equilibrium.[6]

The aspiration level can be thought of as a psychological reference point, as opposed to the social one considered when modeling inequality aversion: the idea behind the concept proposed in Section 3 is that of utilizing the equilibration between upward and downward impulses which is inherent to the IBE, but replacing the aspiration level associated to own-payoff considerations only with equity considerations related to the distance between own and opponent's payoff. The motivation follows from the realization that in non-constant sum games (considered here) subjects' behavior also reflects considerations of equity. In fact, while finite repetition alone has been shown to have limited effectiveness in enlarging the scope for cooperation or retaliation, non-constant sum games offer some cooperation opportunities, and it seems plausible that fairness motives would play an important role in repeated play of this class of games. A suitable consequence of replacing the aspiration level framework with the inequality aversion one is that the original payoffs can be utilized (and should, in order to avoid mixing social

[6]When the IBE is applied to the payoffs belonging to the games truly played by the participants, the gains in fit of the concept over the Nash equilibrium appear to be significantly reduced, indicating that its explanatory superiority depends to a large extent on the payoff transformation, which is itself dependent on the choice of the aspiration level (the pure strategy maximin payoff) and the double weight assigned to losses relative to gains.

and psychological reference points).

Before introducing the other-regarding stationary concepts explored in the next two sections, it is useful to take a closer look at the experiments on the basis of which they will be tested. Table 5 in the Appendix shows the 12 games, 6 constant sum games and 6 non-constant sum games on which Selten and Chmura (2008) have run experiments, which have taken place with 12 independent subject groups for each constant sum game and with 6 independent subject groups for each non-constant sum game. Each independent subject group consists of four players 1 and four players 2 interacting anonymously in fixed roles over 200 periods with random matching. In summary:

Players: $I = \{1, 2\}$
Action space: $\{U, D\}\, x\, \{L, R\}$
Estimated choice probabilities in mixed strategy: $\{P_u, 1 - P_u\}$ and $\{Q_l, 1 - Q_l\}$
Sample size: (54 sessions) x (16 subjects) = 864
Time periods: T=200

In Table 5, a non-constant sum game next to a constant sum game has the same best reply structure (characterized by the Nash equilibrium choice probabilities P_u, Q_l) and is derived from the paired constant sum game by adding the same constant to player 1's payoff in the column for R and to player 2's payoff in the row for U. Games identified by a smaller number have more extreme parameter values than games identified by a higher number; for example, Game 1 and its paired non-constant sum Game 7 are near the border of the parameter space ($P_u \simeq 0.1$ and $Q_l \simeq 0.9$), while Game 6 and its paired non-constant sum Game 12 are near the middle of the parameter space ($P_u \simeq 0.5$ and $Q_l \simeq 0.6$). As pointed out above, IBE involves a transition from the original game to the transformed game, in which losses with respect to the aspiration level get twice the weight as gains above this level. The impulse balance equilibrium depends on the best reply structure of this modified game, which is generally different from that of the original game, resulting therefore in different predictions for the games in a pair.

The present paper utilizes the data on the experiments involving 6 independent subject groups for each of the 6 non-constant sum games (games 7 through 12 in Table 5). As previously anticipated, this class of games is conceptually suitable to the application of the inequality aversion framework. Further, in completely mixed 2x2 games, mixed equilibrium is the unambiguous game theoretic prediction when they are played as non-cooperative one-shot games. Since non-constant sum games provide incentives for cooperation, such attempts to cooperation may have influenced the observed relative frequencies in the experiment by Selten and Chmura (2008). Along these lines, it is particularly relevant to see whether inequality aversion payoff modifications can help improve the fit with respect to these frequencies.

The application of inequality aversion parameters to the impulse balance equilibrium provides an opportunity for testing the fairness model by Fehr and Schmidt (1999) in conjunction with the latter, which is itself a simple yet powerful concept which has proven to be empirically successful in fitting the data in different categories of games while nevertheless being parsimonious (see footnote 12 for remarks on the not fully parameter-free nature of IBE). By including a fairness dimension to it, the hope is to supply favorable empirical evidence and provide further stimulus to expand the types of games empirically tested. Formally, this involves first modifying the payoff matrices of each game in order to account for the inequality parameters (β,α), than creating the impulse matrix based on which the probabilities are computed. In order to clarify the difference between the reference point utilized in Selten and Chmura (2008) (the aspiration level) and that utilized in this paper, it is useful to start by summarizing the mechanics behind the computation of the original version of the IBE. Let's consider the normal form game depicted in Figure 1.

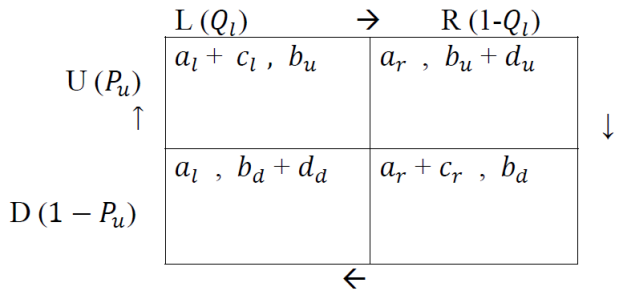


Figure 1: Structure of the 2x2 games (arrows point in the direction of best replies; probabilities in parentheses)

In it, $a_l, a_r, b_u, b_d \geq 0$ and $c_l, c_r, d_u, d_d > 0$. c_l and c_r are player 1's payoffs in favor of U, D while d_u, d_d are player 2's payoffs in favour of L, R respectively. Note that player 1 can secure the highest of a_l, a_r by choosing one of his pure strategies, since if player 1 chooses U, player 2 will certainly choose R as $b_u + d_u > b_u$, while if player 1 selects D, player 2 will opt for L as $b_d + d_d > b_d$. Similarly, player 2 can secure the highest one between b_u and b_d. Therefore, the authors define the aspiration levels for the two players as given by:

$$s_i = \Big\{ max(a_l, a_r),\ for\ i = 1 \quad max(b_u, b_d),\ for\ i = 2 \tag{2}$$

The transformed game (henceforth TG) is constructed as follows: player i's payoff is left unchanged if it is less or equal to s_i, while payoffs in excess of s_i are reduced by half such surplus. Algebraically, calling $x_i^{o,r}$ and $\hat{x}_i^{o,r}$ the payoffs for player i when utilizing own strategy o against rival strategy r, before and after the transformation respectively, the following payoff transformation obtains:

$$\hat{x}_i^{o,r} = x_i^{o,r} - 1/2\, max(x_i^{o,r} - s_i, 0) \tag{3}$$

If after the play, player i could have obtained a higher payoff by employing the other strategy, player i receives an impulse in the direction of the other strategy, of the size of the foregone payoff in the TG. Below, a matrix showing the impulses in the direction of the unselected strategy is given,

based on the game transformation resulting from equation (3):

	L (Q_l)	R (1-Q_l)
U (P_u)	0 , d_u*	c_r* , 0
D ($1 - P_u$)	c_l* , 0	0 , d_d*

Figure 2: Impulses in T.G. in the direction of unselected strategy (probabilities in parentheses)

In Figure 2, d_u^*, c_r^*, c_l^* and d_d^* are the impulses in the direction of the unselected strategy, which are positive whenever the payoff for the alternative strategy was higher than the one obtained with the chosen one. The stars are used to remind the reader that the impulses have size equal to that of the forgone payoff *in the transformed game*, as given by applying equation (3) to the entries of Figure 1, rather than having a magnitude equal to the forgone payoff in the original game (where the payoff differences are given by d_u, c_r, c_l and d_d). The concept of impulse balance equilibrium requires that player one's expected impulse from U to D is equal to the expected impulse from D to U; likewise, player two's expected impulse from L to R must equal the impulse from R to L. Formally,

$$P_u Q_r c_r^* = P_d Q_l c_l^*$$

$$P_u Q_l d_u^* = P_d Q_r d_d^*$$

Which, after some manipulation, can be shown to lead to the following *formulae* for probabilities:

$$P_u = \frac{\sqrt{c_l^*/c_r^*}}{\sqrt{c_l^*/c_r^*} + \sqrt{d_u^*/d_d^*}}; \; Q_l = \frac{1}{1 + \sqrt{\frac{c_l^*}{c_r^*}\frac{d_u^*}{d_d^*}}} \tag{4}$$

3 A model with inequality aversion and regretful behavior

3.1 Mechanics of the Equity-driven IBE

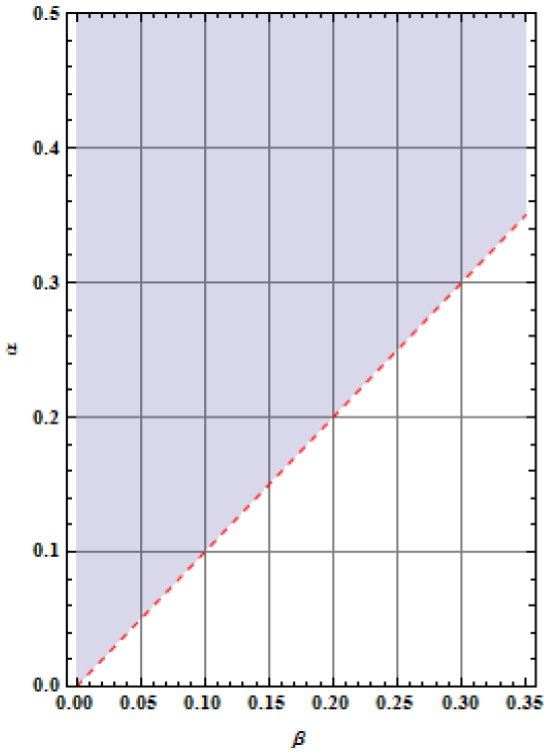
In this section we present a model where "irrational behavior" (i.e. departures from the predictions of the Nash equilibrium) is guided by regret considerations as well as concerns for equity as signaled by relative earnings. In particular, in what follows we will retain the impulse equilibration mechanism, i.e. we will continue to assume that individuals adjust their strategies based on differences between realized payoffs and payoffs obtainable with the alternative strategy, in such a way that in equilibrium each player's upward and downward impulses are equal.

The first departure from IBE will be that we will dispense with two assumptions implicit in the impulse computation presented above, which requires the payoff transformation (and therefore the choice of the aspiration level as the maximin payoff and the choice of a weight equal to 2 to be assigned to losses). Rather, we will stick to the original payoff matrices and consider the impulses to simply be the size of the actual forgone payoffs in the considered game, as given by d_u, c_r, c_l and d_d. While this choice implies a reduction in the concept performance as evidenced by Selten and Chmura (2008) in Figure 11 (page 959), concerning the 12 games they utilize (of which we use the 6 non-constant sum ones), we believe that other reference considerations may play an important role in determining individuals' behavior, and in order to avoid to build an overparametrized model, we have discarded the behavioral assumptions implicit in the payoff transformation. This approach has the advantage of reducing the cognitive burden on the players' part, since they no longer are assumed to assess their performance relative to the maximin payoff in each move, rendering the concept more justified as a bounded rationality one. This translates to replacing (4) with

the following *formulae*, in order to compute the probabilities of play:

$$P_u = \frac{\sqrt{c_l/c_r}}{\sqrt{c_l/c_r} + \sqrt{d_u/d_d}};\ Q_l = \frac{1}{1 + \sqrt{\frac{c_l}{c_r}\frac{d_u}{d_d}}} \tag{5}$$

The second departure from IBE concerning the concept proposed in this section has to do with the introduction of other-regarding distributive concerns, which are taken to affect individuals' subjective utilities. This is done by replacing the aspiration level framework with the inequality aversion one, and doesn't require the computation of the TGs based on aspiration level framing; rather, the original payoffs are now modified by including the inequality parameters (β,α). A cutout of the relevant parameter space (for the games considered here) is described by the highlighted area in Figure 3.



$$\beta_i \leq \alpha_i \text{ and } 0 \leq \beta_i \leq 1$$

Figure 3: A cutout of the correspondence between β_i and α_i (grey area) under the inequity aversion restrictions

Formally, making the perceived payoffs dependent on fairness considerations can be done as follows: recalling that the payoff perceived by an inequity averse individual is affected by his relative standing as given by $U_{ij} = x_i - \alpha_i \, max\left\{x_j - x_i, 0\right\} - \beta_i \, max\left\{x_i - x_j, 0\right\}$, one can modify the matrix in Figure 1 to account for the other-regarding (distributive) reference considerations embodied in the inequity aversion. Table 1, below, contains the proposed payoff modifications:

	L	R
U	$a_l + c_l - \alpha max\{b_u - a_l - c_l, 0\} - \beta max\{a_l + c_l - b_u, 0\}$, $b_u - \alpha max\{a_l + c_l - b_u, 0\} - \beta max\{b_u - a_l - c_l, 0\}$	$a_r - \alpha max\{b_u + d_u - a_r, 0\} - \beta max\{a_r - b_u - d_u, 0\}$, $b_u + d_u - \alpha max\{a_r - b_u - d_u, 0\} - \beta max\{b_u + d_u - a_r, 0\}$
D	$a_l - \alpha max\{b_d + d_d - a_l, 0\} - \beta max\{-b_d - d_d + a_l, 0\}$, $b_d + d_d - \alpha max\{a_l - b_d - d_d, 0\} - \beta max\{-b_d - d_d + a_l, 0\}$	$a_r + c_r - \alpha max\{b_d - a_r - c_r, 0\} - \beta max\{a_r + c_r - b_d, 0\}$, $b_d - \alpha max\{a_r + c_r - b_d, 0\} - \beta max\{b_d - a_r - c_r, 0\}$

Table 1: structure of the 2x2 games accounting for inequality aversion

Note that Table 1 is based on the direct application of the inequality aversion parameters to the payoffs in Figure 1, without making use of the self-centered (psychological) reference point represented by the aspiration level, and given by (2) and (3).The impulses in the direction of the more profitable strategy are now dependent on the objective payoff difference arising from the original matrix *and* on the difference in subjective disutility from inequity aversion associated with the different moves. For example, consider the impulse from D to U for player 1. In the absence of inequity aversion, that is for $\alpha = \beta = 0$, player 1 would experience an upward impulse of size c_l (in place of c_l^* experienced in standard IBE). However, for nonzero inequity aversion parameters, the impulse will be given by $c_l^* = c_l - \alpha \, max\left\{b_u - a_l - c_l, 0\right\} - \beta \, max\left\{a_l + c_l - b_u, 0\right\} + \alpha \, max\left\{b_d + d_d - a_l, 0\right\} + \beta \, max\left\{-b_d - d_d + a_l, 0\right\}$.

It is apparent that this quantity can be larger or smaller than the objective payoff difference c_l, depending on the relative size of the disutility due to inequity aversion.Similarly, now we have the following upward, rightward and downward impulses, respectively: $d_u^* = d_u - \alpha \, max\left\{a_r - b_u - d_u, 0\right\} -$

$\beta\, max\,\{b_u + d_u - a_l, 0\} + \alpha\, max\,\{a_l + c_l - b_u, 0\} + \beta\, max\,\{b_u - a_l - c_l, 0\}\,; c_r^* = c_r - \alpha\, max\,\{b_d - a_r - c_r, 0\} - \beta\, max\,\{-b_d + a_r + c_r, 0\} + \alpha\, max\,\{b_d - a_r - c_r, 0\} + \beta\, max\,\{-b_d + a_r + c_r, 0\}\,; d_d^* = d_d - \alpha\, max\,\{a_l - b_d - d_d, 0\} - \beta\, max\,\{-b_d - d_d + a_l, 0\} + \alpha\, max\,\{-b_d + a_r + c_r, 0\} + \beta\, max\,\{b_d - a_r - c_r, 0\}$

Based on these impulses, and recalling that in equilibrium player i's expected impulse from one of her strategies towards the other pure strategy must be equal to the expected impulse in the opposite direction, the artificial probabilities in (4) can be computed in order to find the mixed strategy equilibrium predictions corresponding to specific values of β and α. Notice that the payoffs in Table (1) and the above impulses are calculated utilizing parameters without indices ($\alpha_1 = \alpha_2$ and $\beta_1 = \beta_2$), that is we assume that all row players and all column players share the same inequity aversion parameters. By doing so, we hope to obtain a parsimonious yet realistic model, whose performance does not rely on the abundance of free parameters; moreover, we believe it important to come up with estimates for the envy and guilt parameters that can be interpreted and confronted with those obtained in other contributions. Such a task would become less transparent without these restrictions. The preceding analysis served to familiarize us to the mechanics behind the first of the two concepts advanced in this paper, namely the equity-driven impulse balance equilibrium. We are now ready to assess the descriptive and predictive success of the original impulse balance equilibrium in comparison to EIBE.

3.2 The first measure of the relative performance of EIBE: best fit

Following a methodology which has been broadly utilized in the literature to measure the adaptive and predictive success of a point in a Euclidean space, the mean squared distance (MSD) of observed and theoretical values is employed.[7] More precisely, let's first focus on the ability of EIBE to describe the choices of a population playing entirely mixed 2x2 games: for each of the

[7] Cf. Erev and Roth (1998), Selten (1991), as well asMarchiori and Warglien (2008) for supporting arguments on the suitability of MSD as a measure of the distance between a model's prediction and the experimental data.

six non-constant sum games considered, a grid search with a mean squared deviation criterion on the (β,α) parameter space has been conducted to estimate the best fitting parameters, that is those that minimize the distance between the data generated by the model and the observed relative frequencies of play.

With this definition in mind, we say that the best overall fit is given by the parameter configuration that minimizes the mean over all games of the distance between the experimental data and the artificial predictions generated by the model. This amounts to first computing the mean squared deviations independently for each game i based on the parameters (β_i, α_i) and then finding the $(\beta, \alpha)_{best\,fit}$ that minimize the average across all games. Algebraically, letting f and p be the N-length vectors of observed and estimated choice frequencies, respectively, we seek to minimize:

$$MSD = \frac{1}{N}\sum_{i=1}^{N} MSD_i \tag{6}$$

where MSD_i is the average of game i's squared distances, given by:

$$MSD_i = \frac{(f_{ui} - P_{ui})^2 + (f_{li} - Q_{li})^2}{2} \tag{7}$$

and f_{ui} and f_{li} are the observed frequencies of playing up and left in game i, respectively, while P_{ui} and Q_{li} are the estimated relative choice probabilities in mixed strategy. Note that a smaller MSD indicates better fit, i.e. a smaller distance to the experimental data. Table 5 in the Appendix and Table 2 present complementary results on the relative performances of the examined stationary equilibrium concepts. In Table 5, in addition to the recorded choice frequencies and Nash equilibrium (NE) predictions, a summary of the results of the explanatory power of EIBE relative to IBE is shown for each non-constant sum game, utilizing both the transformed (TG) as well as the original payoffs (OG). The comparisons between the two concepts are made both within game class (e.g. by comparing the performance within the class of transformed or original games in column 5), and across game class in the last column (e.g. between the performance of

EIBE using original game i and IBE using transformed game i, i=7,...,12).

The *raison d'être* of the two-fold comparison is that not only it is meaningful to assess whether the proposed model can better approximate the observed frequencies than impulse balance equilibrium can, but it is especially important to answer the question: does EIBE outperform IBE when the former is applied to the original payoffs of game i and the latter is applied to the corresponding transformed payoffs? In other words, since the inequality aversion concept overlaps to a certain extent to that of having impulses in the direction of the strategy not chosen, applying the inequality aversion adjustment to payoffs that have already been transformed to account for the aspiration level will result in "double counting".[8] It is therefore more relevant to compare the best fit of EIBE on OG (see rows highlighted in blue in the last column of Table 2) to that obtained by applying impulse balance equilibrium to TG.

Inspection of Table 5 suggests a strong positive answer to the following two relevant questions regarding the ability of the proposed concept to fit the observed frequencies of play: within the same class of payoffs (TG or OG), is the descriptive power of EIBE superior to that of the IBE? And, perhaps more importantly, is this still true when the two concepts are applied to their natural payoff matrices, namely the original and the transformed one respectively? The last two columns of Table II show that, based on a comparison of the mean squared deviations of the predicted probabilities from the observed frequencies under the two methods, the EIBE fares better than IBE when the inequity aversion parameters are fit to each game separately. This result, however, may owe, at least in part, to the fact that a parametric concept, such as the one advanced here (as well as equity-driven QRE introduced in Section 4), is compared to a parameter-free one.

[8]See TG7 and TG12 in Table5 for instances where the best fit is achieved when both inequity parameters are 0 (in contrast to the paired original games, which have nonnegative parameters). Moreover, $(\beta,\alpha)_{TG}<(\beta,\alpha)_{OG}$ for all games, indicating that aspiration level and inequity aversion reference dependence overlap to some extent.

3.3 The second measure of the relative performance of EIBE: predictive power

In order to correct for this advantage, results for the proposed parametric concepts are also reported avoiding to fit them for each game separately. This is done in two ways (as will be further explained below): by utilizing the two parameters that best perform *on all games* in order to derive each game's predictions (and MSD), or by making out-of-sample predictions for each game based on the two free parameters that minimize the MSD of the remaining 5 games.

Let's take a closer look at the evaluation of the performance of equity-driven impulse balance equilibrium concept by means of an assessment of its predictive power. As mentioned, this is accomplished by partitioning the data into subsets, and simulating each experiment using parameters estimated from the other experiments. By generating the MSD statistic repeatedly on the data set leaving one data value out each time, a mean estimate is found making it possible to evaluate the predictive power of the model. In other words, the behavior in each of the 6 non-constant sum games is predicted without using that game's data, but using the data of the other 5 games to estimate the probabilities of playing up and down. By this cross-prediction technique, one can evaluate the stability of the parameter estimates, which shouldn't be substantially affected by the removal of any one game from the sample.[9]

Erev and Roth (1998) based their conclusions on the predictive success and stability of their learning models by means of this procedure, as well as, more recently, Marchiori and Warglien (2008). Table 2 shows summary MSD scores ($100{*}Mean - squared\ Deviation$) organized as follows: each of the first six columns represents one non-constant sum game, while the last column gives the average MSD over all games, which is a summary statistic by which the models can be roughly compared.[10] The first three rows present the MSDs of the NE and IBE predictions (for $\beta = 0 = \alpha$) on

[9] Cross-validation (also known as jackknifing) is extensively discussed inBusemeyer and Wang (2000).

[10] Note that here we restrict attention to the OGs when considering EIBE.

the transformed and original payoffs respectively. The remaining three rows display MSDs of the EIBE model on the original payoffs: in the fourth row, the parameters are separately estimated for each game (12 parameters in total); in the fifth row, the estimated 2 parameters that best fit the data over all six games (and over all but Game 7, the reason will be discussed below), are employed (the same two β,α that minimize the average score over all games are used to compute the MSDs for each game); in the last row the accuracy of the prediction of the hybrid model is showed when behavior in each of the six games is predicted based on the two parameters that best fit the other five games (and excluding Game 7).

Model		**G 7**	**G 8**	**G 9**	**G 10**	**G 11**	**G 12**	**Mean** (s.d.)
NE (on OG)								
0 parameters	All games	6.08	1.23	.354	.708	.422	.064	**1.48** (2.29)
	G8-12							**.555** (.440)
IBE (on OG)								
0 parameters	All games	.330	1.17	1.83	.878	.497	.209	**.819** (.610)
	G8-12							**.917** (.627)
IBE (on TG)								
2 parameters	All games	.315	.035	.416	.224	.094	.205	**.215** (.140)
	G8-12							**.195** (.134)
EIBE by game (on OG)								
12 parameters	All games	.090	.003	.031	.033	.056	.000	**.035** (.034)
	G8-12							**.025** (.020)
6 par. ($\boldsymbol{\beta_i = 0}$)	All games							**.058** (.050)
EIBE best fit (on OG)								
2 parameters								
(.157,.157)	All games	.746	.178	.428	.152	.140	.030	**.279** (.254)
(.253,.259)	G8-12	-	.042	.098	.033	.173	.034	**.076** (.060)
EIBE predict (on OG)								
2 parameters	All games	2.22	.238	.585	.186	.141	.031	**.567 (.837)**
	G8-12	-	.044	.149	.033	.189	.035	**.09** (.074)

Table 2: MSD scores of the considered equilibrium concepts (standard deviations for the means in parentheses)

Table2 summarizes further evidence in favor of the newly developed equity-driven impulse balance equilibrium. One can see from the third row that (as already signaled by Table 5), if the parameters of inequality aversion

are allowed to be fit separately in each game, the improvements in terms of reduction of MSD are significant, both with respect to the Nash and impulse balance equilibrium. In order to consider a more parsimonious version of the model evaluated in this section, the aggregate MSD score of a 1-parameter adaptation of EIBE, which one may call envy-driven IBE, is also reported in the fourth row of Table 2. Note that the overall reduction in the number of parameters from 12 to 6 doesn't come at a dear price in terms of MSD, which goes from 0.35 for the full model to 0.58 for the reduced one, signaling the relative importance of the disadvantageous inequity aversion with respect to advantageous inequity aversion.

Let's now restrict the number of parameters to two (common to all players in all games, cf. row 5 "EIBE best fit" in the above table): the mean MSD is still more than five times smaller than Nash's. If one doesn't include the extremely high MSD reported in both cases for Game 7 (for reasons discussed below), the gap actually increases, as the EIBE's MSD becomes more than seven times smaller than Nash's. With respect to the overall MSD mean of the IBE, when considering all games the proposed concept has a higher MSD, although a similar order of magnitude (.279 and .215 respectively). If one focuses only on games 8-12, again we have a marked superiority of equity-driven IBE over conventional IBE, as the MSD of the latter is more than twice that of the new concept. A similar pattern appears in the last row of the table, concerning the predictive capability: if Game 7 is excluded, the values are in line with the ones obtained in the fifth row, indicating stability of the parameters who survive the cross-validation test. One comforting consideration regarding the appropriateness of the exclusion of Game 7 comes from the widespread anomalous high level of its MSD score in all rows of the table, which for both Nash and EIBE predict is about four times the corresponding mean level obtained over the six games. It is plausible that this evidence is related to the location of Game 7 in the parameter space. It is in fact located near the border, as previously pointed out, and therefore may be subject to the overvaluation of extreme probabilities by the subjects due to overweighting of small probabilities.

The next section considers incorporating fairness motives in the quantal

response equilibrium notion, one that has recently attracted considerable attention thanks to its ability to rationalize behavior observed in experimental games. In addition to providing an interesting case for comparison, it should also allow to shed light on the suspected anomalous nature of Game 7.

4 A model with inequality aversion and noisy behavior

4.1 Mechanics of the Equity-driven QRE

Here we propose an alternative model which shares the aim of the one described in the previous section, namely of accounting for multiple facets that determine individual behavior, but focuses on the role of unobserved factors and stochasticity, in addition to fairness motives (again in the form of inequity aversion). That is, we want to see whether the departures from rational self-regarding behavior observed in the data (as shown by the poor performance of the Nash equilibrium in Table 2 and Table 5), can be accounted for by means of bounded rationality, in the form of stochastic choice, and concerns for relative standing. We assume that, while players attempt to best respond to the opponent's action, they "drift away" due to a preference for equitable earnings on the one hand, and noise in decision making stemming from cognitive limitations or to the presence of unobserved factors rendering behavior more unpredictable on the other hand.

The above is achieved by utilizing the logit version of the quantal response equilibrium concept in conjunction with preferences that are again allowed to be affected by the counterparty's fate, via the inequity aversion parameters. The resulting model is called EQRE. Before showing the results, which are given in Table5 and Table 3 and show an even better overall performance of this concept compared to the one examined above, let's briefly describe the QRE. This probabilistic choice model was introduced by McKelvey and Palfrey (1995), and concerns games with noisy players that base their choices on quantal best responses to the behavior of the other parties, so that deviations from optimal decisions are negatively correlated with the

associated costs. That is to say, individuals are more likely to select better choices than worse choices, but do not necessarily succeed in selecting the very best choice. In the exponential form of quantal response equilibrium, considered here, the probabilities are proportional to an exponential with the expected payoff multiplied by the logit precision parameter (λ) in the exponent: as λ increases, the response functions become more responsive to payoff differences. Formally,

$$P_{ij} = \frac{e^{\lambda \pi ij(P_{-i})}}{e^{\lambda \pi ij(P_{-i})} + e^{\lambda \pi ik(P_{-i})}} \tag{8}$$

Where $i = 1, 2$ stands for the player, P_{ij} is the probability of player i choosing strategy j and π_{ij} is player i's expected payoff when choosing strategy $j \neq k$, given the other player is playing according to the probability distribution P_{-i}.

We move from the above model of stochastic choice where players imprecisely attempt to act rationally and selfishly, to one that, while continuing to postulate noisy behavior, allows it to also respond to equity considerations. This coupling of (imperfect) maximizing behavior and distributive concerns is achieved by replacing the monetary payoffs in 8 with the ones in Table 1, which are reduced in order to account for subjects' resistance to inequitable outcomes as described in 1.

While writing the paper we have become aware that a similar exercise has been performed by Goeree (2000), who successfully employs a model of inequality aversion together with a logit equilibrium analysis in order to explain behavior in experimental alternating-offer bargaining games. One salient difference concerning the two models pertains to the parameterization (see also the related discussion concerning EIBE in Section 3): while here, for the sake of parsimony, we restrict both alpha and beta to be the same for both populations of players (those playing as player 1 and those in the role of player 2), Goeree uses a 4-parameter specification allowing the proposers to have different "guilt" parameter beta from the responders. The 3-parameter specification employed here (i.e. the utility and error parameters β,α and $\lambda$are common to all players), while inevitably resulting in a reduced fit to

the data, is taken with the aim of preserving parsimony and comparability with past and future efforts. In particular, given the payoff structure of the games considered here (which is impartial with respect to the identity of the players), it doesn't seem justified to consider different parameter values for the two populations of students.

4.2 Two measures of relative performance of EQRE: best fit and predictive power

Table 5 in the Appendix is a companion table to Table 5, as it reports the results of comparisons between the model of noisy behavior affected by equity considerations and the standard IBE model employing the aspiration level (and thus the TG); these comparisons are in favor of the former, which outperforms the latter model in each game in terms of smaller MSD. Notice that the penultimate column now compares the performance of the two proposed concepts, showing that EQRE outperforms EIBE in five of the six games.[11]

As before, in order to assess the performance of the concepts over multiple games, the parameters are restricted to be the same over all the games, as shown in the penultimate row in Table 3: EQRE displays a better fit than EIBE (smaller mean square deviation) in all but game 11, achieving a mean MSD of .147 as opposed to .279 for the latter. As for the predictive power, measured for each game by fitting parameters estimated on the remaining five, when all games are considered the mean MSD is substantially lower for the equity-driven QRE, averaging .214 vs. a score of .567 for the equity-driven IBE. Table 3, below, summarizes these comparisons:

Two important considerations should be remarked at this point. Firstly, for what concerns the overall fit, even without excluding the potentially problematic game 7, the EQRE concept outperforms the conventional impulse balance equilibrium applied to the transformed games (MSD scores are .147 and .215, respectively); this is noteworthy, since it wasn't the case for the other hybrid concept.[12] Secondly, the above considerations are con-

[11] in game 12 they achieve a substantially equal equilibrium prediction.

[12] In fact, the impulse balance equilibrium obtains dramatically higher MSD scores when

Model	G 7	G 8	G 9	G 10	G 11	G 12	**Mean** (s.d.)
NE (on OG) 0 parameters	6.08	1.23	.354	.708	.422	.064	**1.48** (2.29)
IBE (on OG) 0 parameters	.330	1.17	1.83	.878	.497	.209	**.819** (.610)
IBE (on TG) 2 parameters	.315	.035	.416	.224	.094	.205	**.215** (.140)
EQRE by game (on OG) 18 parameters	$5.5^{*}10^{-6}$	$2.4^{*}10^{-7}$	$7.5^{*}10^{-6}$	$6.4^{*}10^{-7}$	$7.4^{*}10^{-8}$	$5.7^{*}10^{-6}$	$\mathbf{3.3^{*}10^{-6}}$ ($3.0^{*}10^{-6}$)
Parametric best fit (OG) EIBE ($\boldsymbol{\beta}$=$\boldsymbol{\alpha}$=.16) EQRE ($\boldsymbol{\beta}$=.15,$\boldsymbol{\alpha}$=.24, λ=.43)	.746 .251	.178 .012	.428 .397	.152 .036	.140 .163	.030 .027	**.279** (.279) **.147** (.154)
EIBE vs. EQRE predict (OG) 2 par. EIBE 3 par. EQRE	2.22 .558	.238 .023	.585 .420	.186 .062	.141 .189	.031 .030	**.567** (.831) **.214** (.226)

Table 3: MSD scores of the considered equilibrium concepts

firmed by the predictions obtained with the jackknifing technique: for the EQRE specification the mean MSD score based on cross-predictions is not substantially higher than the one calculated when the parameters that best fit all games are employed (.214 and .147, respectively). This doesn't hold for the EIBE concept, whose score in the prediction field in the last row is roughly double the one in the best fit row (.567 in place of .279). Note also that the average MSD for equity-driven QRE when cross-predicting is approximately equal to the mean score for IBE on all transformed games (.214 for EQRE as opposed to .215 for IBE), further confirming the stability of the parameters in the other-regarding version of QRE. Again, this cannot be said for EIBE, whose score when using parameters fitted out of sample is substantially higher than the score for the 2-parameter impulse balance equilibrium (.567 to be compared to .215).

the original games are employed in place of the transformed ones, with an almost four-fold increase. The intuition behind this is, loosely speaking, that the IBE is not as parameter-free as it looks: that is, by utilizing transformed payoffs for each game (although based on common definition of aspiration level), it effectively allows for game-specific adjustments similar to those obtained by adding a parameter which can take different values in each game.

5 Discussion

This paper is concerned with advancing two empirically sound concepts: equity-driven impulse balance equilibrium and equity-driven quantal response equilibrium: both introduce a distributive reference point to the corresponding established stationary concepts known as impulse balance equilibrium and quantal response equilibrium. The former is modified in order to retain the impulse equilibration due to regret considerations associated with "wrong" plays while discarding the original parameterization (which assigned a double weight to losses with respect to the maximin payoff count, relative to gains), and at the same time build in equity considerations by utilizing the utility functions in 1 in place of the monetary payoffs in Figure 1. Quantal response equilibrium, on the other hand, serves as the basis for a concept that aims at explaining behavior as the result of a mix of rationality, cognitive limitations (these two leading to stochastic best replies to the opponent's action) and fairness motives again in the form of other-regarding inequality aversion. This coupling of (imperfect) maximizing behavior and distributive concerns is achieved by replacing the monetary payoffs in (8) with the ones in Table 1, which are reduced in order to account for subjects' resistance to inequitable outcomes. Before drawing conclusions on the relative performance of the concepts analyzed here, let's take a closer look at the meaning of the two parameters that are common to both equity-driven equilibrium concepts proposed here, and which, consistently with the original specification byFehr and Schmidt (1999), are required to satisfy the constraints $\beta_i \leq \alpha_i$ and $0 \leq \beta_i \leq 1$. As argued in the introduction, this choice is not trivial, and has been taken for the sake of parsimony and comparability. It may, however, be reasonable to extend the standard inequality aversion model in (1) to more general domains accounting for strong altruism as well as spiteful behavior (and is the subject of another ongoing project). In particular, let's consider in turn the implications of relaxing the constraints on β and α, focusing first on the last term of the right-hand side

of (1), representing the positive deviations from the reference outcome (x_j):

$$0 \leq \beta_i \leq 1 \tag{9}$$

Restricting the parameter space to values of β lying between zero and one means, on one hand ($0 \leq \beta_i$), ruling out the existence of spiteful individuals who enjoy being better off than the opponent, and on the other hand ($\beta_i \leq 1$) ruling out the existence of strongly altruistic subjects who care enough about the well being of the other player to incur a decrease in utility which is greater than the payoff difference ($x_i - x_j$). Both possibilities are coherent and some degree of similar pro- and anti-social behavior has been observed in the literature (cf. Bester and Guth (1998), Bolle (2000) and Possajennikov (2000), so excluding them ex ante may bias the analysis against well documented behaviors that appear to have survived the evolutionary pressures shaping the evolution of human preferences. Consider now the second assumption that Fehr and Schmidt make on the parameters, concerning the presumed loss aversion in social comparisons:

$$\beta_i \leq \alpha_i \tag{10}$$

When taken in conjunction with the 'moderate aversion' to advantageous inequality embodied in (9), it seems in fact plausible to postulate that negative deviations from the reference outcome count more than positive ones (disadvantageous inequity induce higher disutility than advantageous inequity). However, when (9) is dropped and agents are free to exhibit strongly altruistic and spiteful behavior, the assumption that β is at most as big as α is no longer justified in all domains. To illustrate this point, let's consider individual i whose preferences satisfy a slight modification of the above parameter restrictions that maintains the asymmetric other regarding preferences of the familiar form.[13] That is, let the parameters modeling other-regarding

[13]The 'conditional altruism' inherent in the inequity aversion framework is preserved so long as α and β are non-negative, implying that both positive and negative deviations from the opponent's outcome induce a utility loss.

behavior satisfy the following inequalities:

$$0 \leq \alpha_i < \beta_i \leq 1 \tag{11}$$

Note that the above inequalities violate (10) while satisfying (11), and still entail that an agent responds with a utility loss to both negative and positive deviations from the reference outcome. The difference lies in β no longer being bounded below α so that its magnitude (representing the altruistic disutility from advantageous inequality) can now be greater than that of the disutility from disadvantageous inequality. Another example of reasonable preferences that are ruled out in the standard inequality aversion model is given by

$$\alpha_i < 0 < \beta_i \tag{12}$$

Loosely speaking, the intuition is that an agent whose preference parameters satisfy the above inequalities simply cares more about the counterparty than about herself, a possibility which may well apply to the truly altruistic agents.[14]

Recent contributions, such as Bolle (2000) and Possajennikov (2000), have drawn the attention on the parameter space concerning the degree of altruism and spite one should allow for when modeling the evolutionary stability of other-regarding preferences. In particular, they have independently criticized and relaxed restrictions that Bester and Güth (1998) had imposed on the parameters. Given the resonance with IA preferences employed here, it is worth briefly introduce some notation from Bester and Güth. Two agents play a symmetric game and are assumed to maximize a weighted sum of the own payoff and of the counterparty's payoff, in order to allow for the possibility that individuals have other-regarding preferences that go beyond their material payoffs. Formally,

$$V_i = U_i(x, y) + \alpha_i U_j(x, y),\ i \neq j \tag{13}$$

[14] In this case, for a given absolute deviation between the two payoffs, the altruistic person will incur a bigger utility reduction when being the one with the higher payoff.

where $U_i(x, y)$ is the material payoff to player i, while α is a preference parameter (subject to evolutionary selection), which is positive under altruism, zero under own profit maximization and negative under spite. As Bolle and Possajennikov show (respectively in the domains of spiteful and altruistic preferences), the preference restrictions imposed by Bester and Güth, namely of ruling out spite and what I will call 'strong altruism', aren't theoretically justified and should be relaxed. More specifically, Bester and Güth assume $0 \leq \alpha \leq 1$ and Bolle and Possajennikov separately show that arbitrarily large negative and positive values of the parameter should be allowed, in order to let the evolutionary pressures ultimately decide whether spite and strong altruism should be ruled out.

With the above discussion in mind, and recalling that the restrictions in (9) and (10) are imposed on the parameter space of both models advanced here despite their restrictive nature, we ask whether the resulting asymmetric inequality aversion significantly contributes to explaining behavior of two populations repeatedly playing six games with random matching.

Based on the comparisons presented in Section 3 and 4 (and in the Appendix), the concept employing the logit equilibrium analysis (and the resulting stochasticity in behavior) on payoffs that are modified to reflect individuals' inequality aversion emerges as the best performing in terms of goodness of fit, among the considered stationary concepts. Following the behavioral stationary concept interpretation of mixed equilibrium, the experimental evidence leads to the conclusion that, among the stationary concepts considered here, the proposed other-regarding generalization of the QRE is the behavioral concept that best models the probability of choosing one of two strategies in various non-constant sum games spanning a wide parameter space.[15]

More specifically, even when restricting the degrees of freedom of the parametric models and comparing the goodness of fit utilizing the same parameters (β,α,λif any) for all six games, the other-regarding QRE outper-

[15]The mixed equilibrium is taken to be the result of evolutionary (or learning) processes in a situation of frequently repeated play with two populations of randomly matched opponents.

forms all of the other stationary concepts considered here.

In summary, the explanatory power of the considered models leads to the following ranking, starting with the most successful in terms of fit to the experimental data (and with the goodness of fit decreasing progressively): EQRE, IBE, EIBE, QRE and Nash equilibrium.[16]

Of course, more parsimonious concepts such as NE, are at a disadvantage when compared to parameterized models such as EIBE and EQRE, due to the parameter-free nature of the former. It should be noted, however, that while Nash equilibrium is truly independent of parameters, the calculation of the impulse balance equilibrium depends on the choice of the aspiration level (the pure strategy maximin payoff) and the double weight assigned to losses relative to gains. In fact, when the IBE is applied to the payoffs belonging to the games truly played by the participants, the gains in fit of the concept over the Nash equilibrium appear to be substantially reduced, indicating that its explanatory superiority depends to a large extent on the payoff transformation. Given the limited theoretical support for the implicit parameters in IBE and that they happen to provide the best fit compared to other choices of aspiration level and loss weight, we have treated them as two free parameters.

Nevertheless, in order to avoid to give an unfair advantage to the proposed parametric models, the ranking presented above is based on rows 1, 3 and 5 in Table 3, which show results obtained avoiding to fit the parameters (if any) to each game separately. It is significant to note that the order of the four concepts established under the above comparison, namely EQRE, IBE, EIBE and NE, is confirmed when restricting attention to the MSD obtained with parameters estimated out-of-sample for the parametric concepts (see the last row of Table 3).

[16] See the grey highlighted rows in Table 3

Appendix

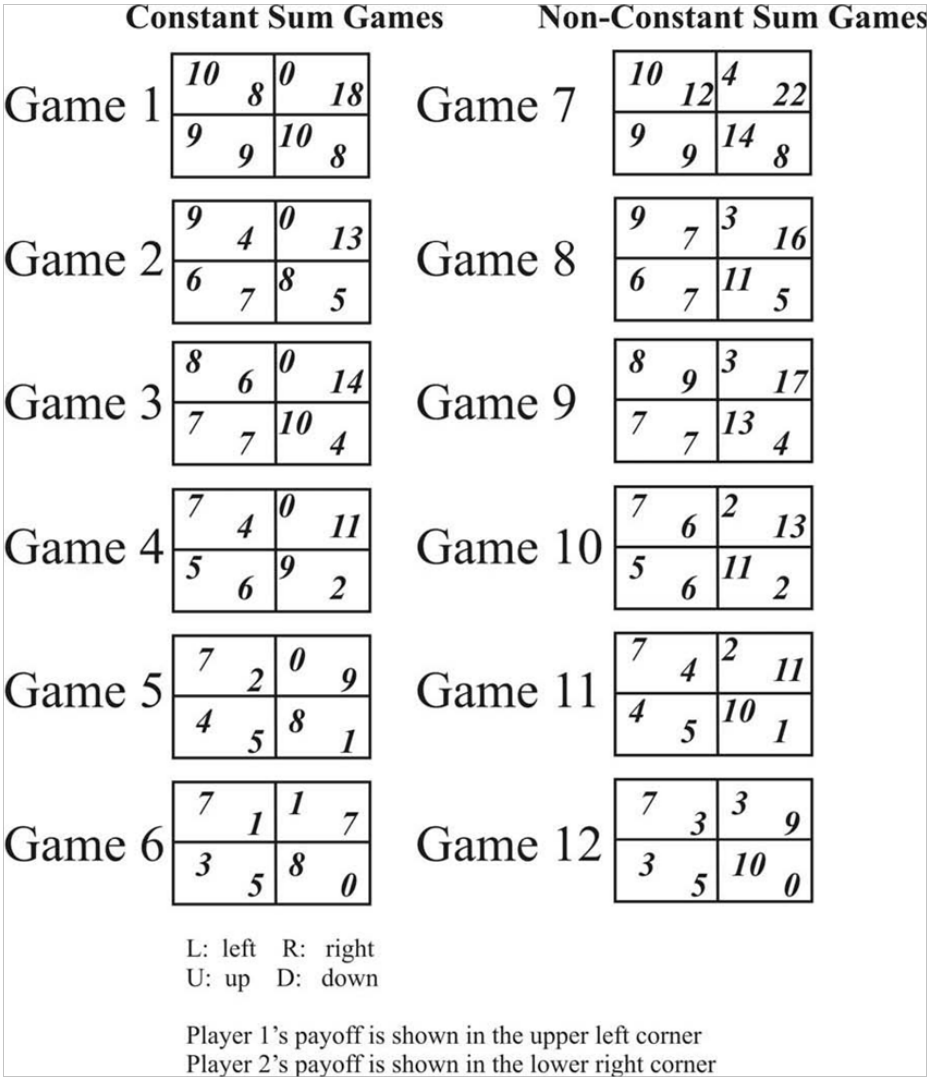


Table 4: Games utilized in Selten & Chmura (2008) and here

	FREQ. f_u, f_l	NE P_u, Q_l	BEST FIT EIBE P_u, Q_l (β, α)	IBE P_u, Q_l $(\beta=\alpha=0)$	MSD EIBE < MSD IBE	MSD EIBE(OG) < MSD IBE(TG)
TG7	.141,.564		.104,.634 (0,0)	.104,.634	NO	*n.a.*
OG7	.141,.564	.091,.909	.099,.568 (.054,.055)	.091,.500	YES	YES
TG8	.250,.586		.270,.586 (.043,.065)	.258,.561	YES	*n.a.*
OG8	.250,.586	.182,.727	.257,.584 (.000,.471)	.224,.435	YES	YES
TG9	.254,.827		.180,.827 (.07,.10)	.188,.764	YES	*n.a.*
OG9	.254,.827	.273,.909	.233,.840 (.330,.330)	.162,.659	YES	YES
TG10	.366,.699		.355,.759 (.089,.134)	.304,.724	YES	*n.a.*
OG10	.366,.699	.364,.818	.348,.717 (.253,.253)	.263,.616	YES	YES
TG11	.331,.652		.357,.652 (.012,.018)	.354,.646	YES	*n.a.*
OG11	.331,.652	.364,.727	.343,.642 (.000,.415)	.316,.552	YES	YES
TG12	.439,.604		.496,0.575 (0,0)	.496,.575	NO	*n.a.*
OG12	.439,.604	.455,.636	.439,.604 (.017,.397)	.408,.547	YES	YES

Table 5: Performance of the proposed concepts with parameters estimated for each game: EIBE vs. IBE

	FREQ. f_u, f_l	NE P_u, Q_l	BEST FIT EQRE P_u, Q_l (β, α, λ)	IBE P_u, Q_l $\beta=\alpha=0$	MSD EQRE < MSD EIBE	MSD EQRE(OG) < MSD IBE(TG)
TG7	.141,.564			.104,.634		*n.a.*
OG7	.141,.564	.091,.909	.141,.564 (.105,.209,.335)	.091,.500	YES	YES
TG8	.250,.586			.258,.561		*n.a.*
OG8	.250,.586	.182,.727	.250,.586 (.059,.431,.310)	.224,.435	YES	YES
TG9	.254,.827			.188,.764		*n.a.*
OG9	.254,.827	.273,.909	.254,.827 (.083,.316,.600)	.162,.659	YES	YES
TG10	.366,.699			.304,.724		*n.a.*
OG10	.366,.699	.364,.818	.366,.699 (.362,.240,.310)	.263,.616	YES	YES
TG11	.331,.652			.354,.646		*n.a.*
OG11	.331,.652	.364,.727	.311,.652 (.003,.02,.910)	.316,.552	YES	YES
TG12	.439,.604			.496,.575		*n.a.*
OG12	.439,.604	.455,.636	.439,.604 (.042,.137,.550)	.408,.547	same	YES

Table 6: Performance of the proposed concepts with parameters estimated for each game: EQRE vs. IBE and EIBE

References

Bester, H., Guth, W., 1998. Is altruism evolutionarily stable? Journal of Economic Behavior & Organization 34 (2), 193–209.

Bolton, G. E., 1991. A comparative model of bargaining: Theory and evidence. American Economic Review 81 (5), 1096–136.

Busemeyer, J. R., Wang, Y., Mar. 2000. Model comparisons and model selections based on generalization criterion methodology. Journal of Mathematical Psychology 44 (1), 171–189.

Colander, D., Holt, R., Rosser, B., 2004. The changing face of mainstream economics. Review of Political Economy 16 (4), 485.

Erev, I., Roth, A. E., 1998. Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria. American Economic Review 88 (4), 848–81.

Falk, A., Fischbacher, U., 2006. A theory of reciprocity. Games and Economic Behavior 54 (2), 293–315.

Fehr, E., Schmidt, K. M., 1999. A theory of fairness, competition, and cooperation. The Quarterly Journal of Economics 114 (3), 817–868.

Goeree, J., 2000. Asymmetric inequality aversion and noisy behavior in alternating-offer bargaining games. European Economic Review 44 (4-6), 1079–1089.

Gowdy, J. M., 2008. Behavioral economics and climate change policy. Journal of Economic Behavior & Organization 68 (3-4), 632–644.

Kahneman, D., Knetsch, J. L., Thaler, R., 1986. Fairness as a constraint on profit seeking: Entitlements in the market. American Economic Review 76 (4), 728–41.

Kahneman, D., Tversky, A., 1979. Prospect theory: An analysis of decision under risk. Econometrica 47 (2), 263–91.

Kirchsteiger, G., 1994. The role of envy in ultimatum games. Journal of Economic Behavior & Organization 25 (3), 373–389.

Levine, D. K., 1998. Modeling altruism and spitefulness in experiment. Review of Economic Dynamics 1 (3), 593–622.

Marchiori, D., Warglien, M., 2008. Predicting human interactive learning by Regret-Driven neural networks. Science 319 (5866), 1111–1113.

McKelvey, R. D., Palfrey, T. R., Jul. 1995. Quantal response equilibria for normal form games. Games and Economic Behavior 10 (1), 6–38.

Possajennikov, A., 2000. On the evolutionary stability of altruistic and spiteful preferences. Journal of Economic Behavior & Organization 42 (1), 125–129.

Rosser, B. J., Cramer, K. L., Holt, R. P. F., Jun. 2010. European Economics at a Crossroads. Edward Elgar Pub.

Selten, R., 1991. Properties of a measure of predictive success. Mathematical Social Sciences 21 (2), 153–167.

Selten, R., Buchta, J., Feb. 1994. Experimental sealed bid first price auctions with directly observed bid functions. Tech. rep., University of Bonn, Germany.

Selten, R., Chmura, T., 2008. Stationary concepts for experimental 2x2-Games. American Economic Review 98 (3), 938–966.

The survival of the conformist: social pressure and renewable resource management

Alessandro Tavoni*, Maja Schlüter†

Abstract

This paper examines the role of pro-social behavior as a mechanism for the establishment and maintenance of cooperation in resource use under variable social and environmental conditions. By coupling resource stock dynamics with social dynamics concerning compliance to a social norm prescribing non-excessive resource extraction in a common pool resource (CPR), we show that when reputational considerations matter and a sufficient level of social stigma affects the violators of a norm, sustainable outcomes are achieved. We find large parameter regions where norm-observing and norm-violating types coexist, and analyze to what extent such coexistence depends on the environment.

Keywords: cooperation, social norm, ostracism, common pool resource, evolutionary game theory, replicator equation, agent-based simulation, coupled socio-resource dynamics

*Advanced School of Economics at the University of Venice, Cannaregio 873, 30121 Venice, Italy. Email: alessandro.tavoni@unive.it

†Leibniz-Institute of Freshwater Ecology and Inland Fisheries, Müggelseedamm 310 12587 Berlin, Germany. Email: schlueter@igb-berlin.de

1 Introduction to Chapter 2

Local and global ecosystems are under growing pressure worldwide, and beyond any doubt their sustainable management cannot be achieved without the stakeholders' cooperative efforts. History has taught us that the livelihood of our species is inextricably related to our ability to cooperate, in the sense of restraining use of natural resources to sustainable levels, rather than giving in to excessive resource appropriation. However, depending on the characteristics of the system at hand, tensions between individual and collective good may undermine such norm of restraint. CPRs where beneficiaries of the resource have open access to it are a notable case of appropriation externality paving the way for short-sighted resource utilization: all agents would be better off if they collectively restrained extraction, but if the impact of one's action on the resource stock is ignored, it is individually rational not to do it. Thus, maintaining cooperation against the myopic self-interest of a potentially large fraction of individual users unwilling to restrict their behavior for the collective good, and despite growing environmental pressure, is a challenging task that often depends on a multitude of factors, as both successful and unsuccessful environmental management has shown. Nevertheless, field work, controlled experiments involving participants playing stylized games aimed at reflecting the trade-offs inherent in these social dilemmas (e.g., CPR and public good games), as well as casual observation, suggest that human beings are able to overcome the obstacles to cooperation in a variety of settings.

Many explanations have been proposed to account for the widely-observed departures from the rational-agent models' predictions of collectively inefficient resource management in the absence of regulatory institutions. Established mechanisms that have been advanced to account for the evolution of cooperation are, following Nowak (2006): kin selection (the inclination of related individuals to engage in cooperative behavior), direct reciprocity (the "I will scratch your back if you scratch mine" attitude towards reciprocating), indirect reciprocity (I will scratch your back because someone else scratched mine), network reciprocity (spatial structure is assumed to allow for unevenly mixed populations where some individuals interact more frequently than others) and multilevel selection (where the population is divided into groups whose members are allowed to enact different strategies depending on whether they are matched with own-group members or with members of other groups).[1]

[1] More recently the term of assortment, indicating the "degree of segregation of different

These mechanisms have been shown to suffice for the evolution of cooperation in Prisoner's Dilemma games whenever the payoffs are such that the benefit-to-cost ratio of the cooperative action exceeds a certain mechanism-specific threshold. While the above mechanisms incorporate some of the empirically observed factors that influence the success of collective action, such as the topology of interactions and group size, we postulate that the link to other important drivers of cooperation needs to be made explicit if one wants to attempt to bridge the gap between the empirical findings on commons management and the theory.

In the present paper we aim to analyze a simple model that departs from the full-rationality paradigm placing emphasis on two such drivers: the presence of individuals with other-regarding preferences and the conformist pressure in the direction of norm compliance arising from fear of community disapproval. Laboratory experiments, such as Fehr and Fischbacher (2002) and Maier-Rigaud et al. (2008), respectively suggest that both are relevant, while contributions from social psychology (Cialdini, 1984) and the empirical literature on the commons stress the importance of the second driver. For what concerns the empirical findings, the work of Ostrom (1990 and 2007) has suggested that many CPRs have escaped the trap of the tragedy of the commons by being managed in a self-organizing manner with mechanisms such as rules, norms and graduated sanctions favoring cooperation among users. Such mechanisms may result from the repeated interactions of the resource users: given suitable conditions, for example in terms of knowledge of the ecosystem by the community members, they could develop a sense of what behavior is acceptable with respect to resource use.

The enforcement of the proper behavior or sanctioning thereof will take different forms depending on the features of the social and resource system at hand. One increasingly studied class concerns social norms. Following the work of Sethi and Somanathan (1996), much attention has been given to the role of costly punishment of defectors in promoting cooperation; recent contributions aimed at extending their setup have been proposed by Noailly et al. (2007) and Sethi and Somanathan (2006) While both retain the three agent types format, with defectors, cooperators and enforcers bearing the costs of punishment, the

types of individuals into different groups" as Pepper (2007) puts it, has gained consensus among scholars for its generality. See van den Bergh and Gowdy (2009) and Fletcher and Doebli (2009) for recent contributions to the group selection debate. It should be noted that many other mechanisms with a certain degree of overlapping characteristics have been employed in various disciplines to highlight the tension between in-group and outsiders; among others, parochialism (Choi and Bowles 2007,) and homophily (Currarini et al. 2009).

former allow for spatial structure in the interactions, and the latter introduce a concern for reciprocity among the agents. Yet, the empirical literature on the commons argues that a variety of sanctioning mechanisms against norm violators are utilized to promote successful management of irrigation systems, fisheries, pastures and forests (Ostrom 1990, Baland and Platteau 1996, ch.8 and 11). Moreover, as suggested by the literature on social capital (Bowles and Gintis 2002, Osés-Eraso and Viladrich-Grau 2007, Iwasa 2009), resource appropriators embedded in a social context can often rely on a wider set of tools than the traditionally considered costly sanctioning of free-riding behavior. When the result of one's actions is observable, be it the resource extraction itself or the outcome of a productive activity which is dependent on the latter, field and experimental evidence suggests that individuals belonging to a community act more cooperatively than when in isolation, as a result of their exposure to social reprobation.[2]

In the present paper we focus on one such mechanism, which we term equity-driven ostracism, by which we refer to the exclusion of norm-violators from community privileges or social acceptance.[3] The underlying idea is that appropriators' decisions about how much effort to exert in the extraction of a natural resource are based on the prevailing norms that have emerged in the community, in addition to the usual efficiency considerations. As a result of the compliance decision with respect to the norm, those who deviate (the defectors) may be refused resources and support by those who comply. As an example, the ostracism costs considered here could be thought of as originating from destruction of defector's crop by the cooperators, or simply from refusal of help by the community towards a defector in the form of denial of loan of machinery or means of transportation needed to take the harvest to the market. The rationale for this behavior is that, in a community of individuals who share access to a natural resource, those who restrain their extraction level to the socially acceptable level will not show the same level of support they have for fellow cooperators, when it comes to defectors. The inherent tradeoff between unrestricted profit seeking and norm adherence can be visualized in Figure 1.

This schematization allows us to highlight the ever-changing conditions faced

[2] According to Gowdy (2008): "Experimental results from behavioral economics, evolutionary game theory and neuroscience have firmly established that human choice is a social, not self-regarding, phenomenon. [...] Human decision-making cannot be accurately predicted without reference to social context". Recent evidence on the importance of the social context in guiding individual behavior is found in Fehr and Fischbacher (2002) and Akerlof (2007).

[3] The term ostracism is to be read here as a refusal of help towards norm violators rather than a physical displacement of the latter, as will be clear from the following discussion.

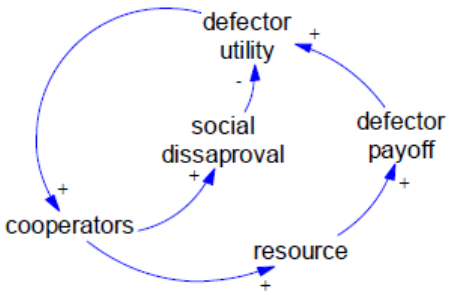


Figure 1: Interactions between the composition of the population, resource abundance and norm enforcement

by appropriators choosing between extraction patterns. If the number of cooperators increases, so will the resource stock, favoring the defectors the most (due to their behavior being unrestricted by the norm) and rendering opportunistic behavior more profitable, thus posing unfavorable selective pressure on the cooperators' population. These trends are captured by the outer arrows originating from the cooperators in Figure 1, with favorable (unfavorable) changes marked with a "+" ("-") sign to the right of the arrows. However, as shown by the inner arrows, an increase in the number of cooperators also leads to greater social disapproval towards norm violators, as fewer defectors will face ostracism by a larger community of norm followers denying them the benefits of cooperation. We implement this simple mechanism in an evolutionary framework in order to allow for departures from optimizing behavior as prescribed by Nash equilibrium. Namely, rather than assuming that the resource appropriators instantaneously maximize their material well-being in a repeated-interaction model with discounted future, we let evolution gradually shape the proportion of agents playing a given strategy by favoring the more successful one. The advantage of this bounded rationality approach (imitate the successful behavior with inertia) is that it avoids the downfalls of the multiplicity of equilibria and lack of robustness to noise, while retaining the behavioral tendency to move in the direction of a more profitable strategy.[4]

The results of the analysis, presented in Section 2, suggest that both monomorphic and dimorphic populations emerge: that is we find stable full compliance and full defection equilibria as in Sethi and Somanathan (1996), but also mixed

[4]For a critique of the commonly used approaches in the economic analysis of common property, see Sethi and Somanathan (2006).

equilibria where both types coexist. In Section 3 we investigate the impact of variation in environmental conditions, by allowing for variability in the rate of resource regeneration as well for a multiplicity of extraction strategies subject to evolutionary pressure. Section 4 provides conclusive remarks.

2 A model of coupled socio and resource dynamics

We examine the role of other-regarding behavior as a mechanism for the establishment and maintenance of cooperation in resource use under variable social and environmental conditions. This is done by modeling the evolution of compliance to a social norm prescribing conformity to an agreed extraction level, and its coevolution with a CPR stock dynamics. The coupled dynamics allows us to investigate the stability of cooperation in a population of resource users who have symmetrical access to it and are not only concerned with own yield from productive use of the resource, but also with their status with respect to other community members, as acceptance to the community is at stake. Pay-off comparisons (e.g. with respect to crop size) lead to ostracism of the norm violators by the cooperating community, which denies defectors the benefits of resource and knowledge support, imposing losses on them that may offset those incurred by cooperators when restricting resource extraction practices to more sustainable ones. That is, individuals face a trade-off: on the one hand they can extract resource at individually optimal (but socially detrimental) levels without restricting usage, or on the other hand they can constraint themselves to a socially agreed upon acceptable level. By doing so, their conventional materialistic pay-off is necessarily below that of the non-cooperating agents, because of the above mentioned lower extraction and consequently reduced yield (which is increasing in the extraction effort). However, violators of the social norm, insofar as they are recognized as such, are penalized by being excluded from the help of the cooperating community (e.g. in bringing the yield to the market): such defectors-specific ostracism costs have a variable magnitude that depends on the relative size of the cooperating community, since at low frequencies of cooperative agent types, the defectors will incur only mild consequences, but at high enough frequencies of cooperators, the former may incur high enough ostracism costs that it will be advantageous to be part of the sustainable community. Lindbeck (1997) suggests this network good property of norms: “a

social norm is felt more strongly, the greater the number of individuals who obey it. Thus, the adherence to a social norm is a choice conditioned on other individuals' adherence to the same norm. The psychological explanation for this type of behavior may be either that disapproval from others is more troubling if expressed by many people than by few or that other people's behavior is assumed to signal information about what is proper or potentially successful behavior." Agents are considered as productive units (one can think of an agent as an individual or a family), whose share of the total production (e.g. the size of their crop) is proportional to the share of their appropriation effort with respect to the aggregate effort. Their source of revenue is assumed to positively depend on two factors: the availability of an indispensable resource for both productivity and livelihood, such as water broadly conceived, and the amount of effort agents put in their productive (income-generating) actions, which is (positively) affected by resource abundance. That is, both the resource and the appropriation effort enter in the agents' (twice-continuously differentiable) production function *f(E,R)*, where E represents the community effort (e.g. the aggregate water usage) resulting from the actions of the n agents comprising it, and R is the resource available to the community (which may either be entirely consumed in a given time period, or saved in part for future consumption). Formally, letting e_i be the individual effort (i.e. his/her resource uptake), which can either take value e_c for a cooperator or e_d for a defector, with $e_c<e_d$ due to the more sustainable practices of the former, the following inequalities are therefore assumed to hold: $\frac{\partial e_i}{\partial R} > 0$, $\frac{\partial f(E,R)}{\partial E} > 0$, $\frac{\partial f(E,R)}{\partial R} > 0$, $\frac{\partial^2 f(E,R)}{\partial E^2} \leq 0$, $\frac{\partial^2 f(E,R)}{\partial E \partial R} \geq 0$, $\frac{\partial (f(E,R)/E)}{\partial R} \geq 0$. [5]Let's go in further detail about the model. It is useful to consider again the joint level of effort E resulting from the actions of the n agents choosing their level of effort e_i; letting $f_c \in [0,1]$ be the proportion of cooperators, we have $E(f_c, R) = f_c * n * e_c(R) + (1 - f_c) * e_d(R)$. We assume that n is fixed, so that entry is ruled out, while f_c is continuous and non-negative, and see that for positive levels of e_c and e_d, the total level of effort is a decreasing function of the frequency of cooperators. The two effort levels, that are here assumed to sum up the behavioral inclinations of all agents in the community, are bounded below by the collectively efficient resource use level and above by the static Nash equilibrium level. This amounts to require

[5]These assumptions are generally employed in the literature concerning resource exploitation in a common pool resource, such as, for example, a fishery where a community of fishermen have access to it and each can decide on the individual level of exploitation (jointly affecting the sustainability of the resource utilization). Cf. Sethi, Somanathan (1996), Xepapadeas (2005) and Oses-Eraso, Viladrich-Grau (2007).

that both agent types follow practices that are above those that would maximize collective utility, but to a different extent: the defectors ignore the emergent social norm prescribing the socially agreed-upon acceptable individual effort e_c by choosing a greater level e_d (up to the individually rational but inefficient Nash equilibrium level resulting in excessive resource use), while cooperators stick to e_c, which, as a special case, may coincide with the level that efficiently trades off the individual incentive towards high or uncoordinated resource utilization with the social need to impose constraints to guarantee a sustainable use (which ultimately benefits the individuals). Letting E_{eff} be the community efficient level, $e_{eff} = E_{eff}/n$ the corresponding individual efficient level under symmetry, and e_{Nash} be the Nash equilibrium level of effort, we formalize what stated above as: $e_{eff} \leq e_c < e_d \leq e_{nash}$.[6]These conditions guarantee that, at the aggregate level, positive rents from productive use of the resource can be maintained. That is, the average product of labor is assured to be above the opportunity cost of labor independently from the share of defectors, providing the incentive for agents to increase their resource use (as they can earn positive profits for positive levels of effort). It is further assumed that $f(0, R) = 0 = f(E, 0)$ for the obvious reason that strictly positive levels of effort and resource are required to generate income via the function $f(E, R)$. The individual payoff given resource R and the behaviour of all community members is:

$$\pi_i(e_1, e_2, \ldots, e_n, R) = p\frac{e_i}{E}f(E, R) - we_i \tag{1}$$

Letting R^* the equilibrium resource level (to be defined more precisely below) and $\Pi(e_1, e_2, \ldots, e_n, R) = \sum_{i=1}^{n} \pi_i = f(E, R) - wE$, the optimal solution to the aggregate payoff maximization problem is given by $E_{eff} = \arg\max_E(\Pi)$, and satisfies $f'(E_{eff}, R^*) = w/p$, where w/p is the ratio between the opportunity cost of labor w and the world price p of the resource-absorbing good produced.[7]

For the sake of compactness and to stress that the payoff to i is only in-

[6]Note that a direct implication of such constraints is that $E_{eff} \leq E \leq E_{nash}$. See Dasgupta and Heal (1979, 55-60) for a comprehensive treatment of exhaustible resources, and Oses-Eraso, Viladrich-Grau (2007, 398) for the description of a process leading to the prevalence of one representative strategy for each type of behavior. In §2.3 we restrict attention to $e_{eff} = e_c$, i.e. we hypothesize that the emergent norm prescribes collectively optimal extraction levels, but, differently from the cited literature, we consider what happens for different magnitudes of defection (ranging from slightly above e_c to Nash effort levels).

[7]An example of a function guaranteeing the existence of an optimal solution, at each point in time, to the aggregate payoff maximization problem, is the familiar Cobb-Douglas formulation with decreasing returns to scale: $f(E, R) = E_t^{\alpha} * R_t^{\beta}$, $\forall E \geq 0, R > 0$ and $\alpha + \beta < 1$.

directly affected by the others' choice of effort (through $f(E, R)$), we will use the notation $\pi_i(e_i, R)$ below. We know that, due to the negative appropriation externality arising from the disconnect between individual extraction and knowledge of its effect on the resource stock, the aggregate level of effort in equilibrium if all agents play according to the Nash equilibrium will be above E_{eff} as each individual will treat the resource stock as fixed and therefore extract more resource than is efficient.

2.1 Resource dynamics

Let's turn to the resource dynamics and its interaction with the social dynamics occurring as a result of human action. Focusing on the ecological features governing the resource first, in the absence of human appropriation we are left with the constant resource inflow (c) and a term dependent on the resource level (R) as well as on three parameters (d, k and R_{max} governing the discharge, curvature and maximum storage capacity) to account for a positive rate of growth up to the R_{max} (e.g. the upper limit in a groundwater aquifer's intake), which becomes negative beyond that level. We follow Ibañez (2004) for what concerns the above mentioned ecological variables, and include the aggregate resource use by the individuals (ER), which appear as the last term of the following equation:

$$\dot{R} = c - d(\frac{R}{R_{max}})^k - ER \tag{2}$$

$\dot{R}$ indicates the time derivative of the resource stock, i.e. its overall rate of change resulting from the interaction of replenishment, discharge and utilization. Note that the resource replenishment rate, which is captured by the first term in the right-hand side of (2), is exogenous with respect to the frequency of cooperators (and consequently the resource extraction), which affects instead the second and last terms, respectively representing the limits to resource accumulation (due to stock effects) and the resource utilized by the community for productive tasks such as irrigation.

For the sake of concreteness, one can think of agents extracting water for irrigation purposes from an underground reservoir subject to replenishment due to snowmelt or rain, whose ability to store water has a natural upper bound (R_{max}). Beyond it, discharge occurs at a rate which is increasing in the deviation from the maximum storage capacity. Two facts are worth noting at this point.

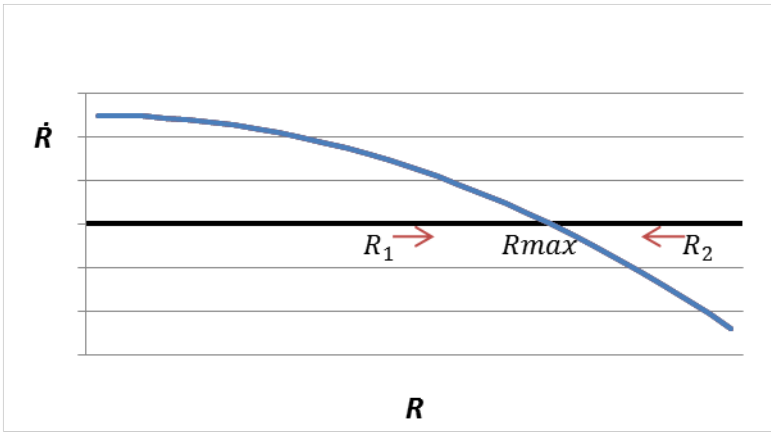


Figure 2: The rate of change of the resource as a function of its stock, in the absence of appropriation

First, in the absence of extraction, the equilibrium resource level will settle on R_{max}. This since, if the stock at one point in time is to its left (R_1 in Figure 2), the resource will continue to accumulate ($\dot{R} > 0$) until R_{max} is reached; if instead the stock at a given time is to its right (R_2), discharge will bring it back to R_{max}. Secondly, due to human extraction, the equilibrium resource level R^* will actually be below the maximum storage capacity.[8] With these notions in mind, we are now ready to shift attention to the strategies and tradeoffs faced by the two types of agents.

2.2 Equity-driven ostracism

Recall that equation (1) represents the amount an individual appropriator can make based on his/her effort and the yield, abstracting from costs originating from social stigma (and the consequential ostracism imposed by the community

[8] In fact, due to the boundary conditions on the effort levels and (1), the equilibrium resource level will satisfy the condition $0 < R_{nash} \leq R^* \leq R_{eff} < R_{max}$, where R_{nash} is the resource level corresponding to monomorphic Nash behavior (a population comprised solely of individuals maximizing profits taking R as exogenous), and R_{eff} is the socially optimal level that would obtain if all individuals jointly maximized collective welfare (effectively internalizing the appropriation externality). Therefore, depending on the population composition, and consequently on the aggregate extraction, the equilibrium level R^* will be closer to one of the above two boundaries: in a dimorphic population comprised of a majority of defectors, R^* will be close to R_{nash}, while its distance from R_{eff} will be shorter the more the number of cooperators. Note that according to the above inequality, even under full defection there is a positive resource value ($0 < R_{nash}$) guaranteeing the assumed positive rents.

on defectors). This amount is proportional to the aggregate payoff (itself a function of the market price of the final yield p and of the production function f), in relation to the individual's resource uptake e_i, which positively enters in the first term in the right hand side of (1) and negatively in the second term representing the work-related costs. To account for the costs accruing to norm violators when identified by the community as such, we incorporate them as shown:

$$U_i = \pi_i - \omega(f_c) * max\left\{\frac{\pi_i - \pi_c}{\pi_d}, 0\right\}$$

$$= e_i(p\frac{f(E,R)}{E} - w) - \omega(f_c) * max\left\{\frac{e_i - e_c}{e_d}(p\frac{f(E,R)}{E} - w), 0\right\} \quad (3)$$

This translates to a payoff to a norm complier (cooperator) which is simply given by

$$U_c = \pi_c \quad (4)$$

while a norm violator will be subject to[9]:

$$U_d = \pi_d - \omega(fc)\frac{\pi_d - \pi_c}{\pi_d} \quad (5)$$

Recalling the tradeoffs highlighted in Figure 1, one sees from the comparison of (4) and (5) that, for what concerns the productivity, defectors have an advantage ($\pi_d > \pi_c$) as a consequence of their higher appropriation; due to stock effects, such productivity advantage positively depends on the relative abundance of cooperators. On the other hand, defectors are deprived of the communitarian social capital and experience a reduction to the income generated with resource-intensive productive activities, while cooperators can also tap in the community for help and thus enjoy the entire yield π_c. As compliance to the social norm is voluntary and observable, these benefits are denied to non-members; further, it is assumed that the community ostracism function $\omega(f_c)$ is increasing and concave in the number of participants (abiding to the social

[9]Raakjær Nielsen and Mathiesen (2003) found, among the five more relevant factors affecting compliance in Danish fisheries: the economic gains to be obtained from noncompliance, deterrence and sanctioning costs, and the presence of "norms (behaviour of other fishers) and morals". In (5), the gains appear in π_d, while the losses due to the sanctions and the comparison with others are captured by the product $\omega(fc)((\pi_d - \pi_c)/\pi_d)$. Notice that the strongest action against defectors will be taken when the number of cooperators is largest (i.e. $\pi_d - \pi_c$ and $\omega(fc)$ are highest), while when defection is spread all over it will go almost unnoticed (i.e.$\pi_d \approx \pi_c$ and $\omega(fc)$ is low).

norm), due to the network good characteristics highlighted at the beginning of this section. Notice also that it multiplies the ratio between the payoff difference and the defector's payoff, to model a reaction by the cooperators which is stronger the larger the relative intensity of the defection (leading to a larger negative productivity gap of norm followers with respect to defectors).

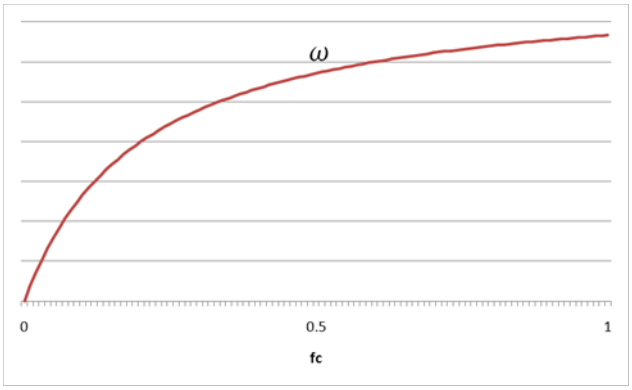


Figure 3: The ostracism function. A higher f_c (larger cooperator community) increases the ostracism's severity. Here we plotted $\omega(f_c) = \delta \frac{f_c}{\lambda + f_c}$, where δ is a scaling factor that will be used in Section 3 to model the community stringency in enforcing the norm

2.3 Evolution of ostracism

The analysis of the behavioral evolution of agents facing decisions on their resource practices is conducted by means of replicator dynamics. By so doing, we avoid the complete rationality requirements typical of models of optimization, while retaining (myopic and lagged) convergence towards better outcomes due to the imitation of successful behavior. Such an approach is particularly well-suited to the analysis of the evolution of norm adoptions as it allows to focus on emergent phenomena without being confined, as is the case for neoclassical analysis, to equilibrium outcomes and representative agents solely described by their optimizing behavior. Here, rather than rationally best responding to the actions of others as in Nash equilibrium, individuals update their strategies when given the option, and switch to the strategy of the agent with which they are randomly matched if the utility of the latter is above the individual's.[10] It can be shown that such strategy revision takes place with a probability which is

[10] See Taylor and Jonker (1978) or Weibull (1995) for details on replicator dynamics.

proportional to the payoff difference with respect to the average: if, for example, the average is well above the payoff of a cooperator, he or she is more likely to notice the benefits from switching than if the average was only slightly above the agent's payoff. Formally, this leads to the 2-strategy replicator dynamics, which combined with (3) yields, after rearranging terms:

$$\dot{f}_c = f_c(U_c - \bar{U}) = f_c(1 - f_c)(U_c - U_d) = f_c(1 - f_c)\frac{\pi_d - \pi_c}{\pi_d}(\omega(f_c) - \pi_d) \quad (6)$$

The dotted superscript stands for time derivative: equation (6) models the evolution of cooperating types. We are interested in the nullclines satisfying $\dot{f}_c = 0$: in addition to the monomorphic outcomes characterized by one type of agent only, we look for solutions in which positive amount of both types coexist (with $f_c \neq 0$ and $f_c \neq 1$). That is,

$$({f_c}^*, R^*) : \theta({f_c}^*, R^*) = \frac{\pi_d(e_d, R^*) - \pi_c(e_c, R^*)}{\pi_d(e_d, R^*)}(\omega({f_c}^*) - \pi_d(e_d, R^*)) = 0 \quad (7)$$

The system described in (2)-(6) can be represented in the (μ, fc) parameter space, where μ is the effort multiplier between a cooperator and a defector; for instance, in Figure 4, $\mu = 2$ signifies that the defector's effort is twice the cooperator's effort, while μ_{nash} signifies that $e_d = e_{nash}$. While only one type of defector is considered at a time (as well as one cooperator type satisfying $\mu = 1$ and extracting at the collectively efficient level), Figure 4 allows one to investigate the prospects for cooperation for different levels of defection. We deem useful to condense this information in the compact graphical tool below, since, depending on the social and ecological characteristics of the system under consideration, the concept of defection may vary greatly. For example, in a relatively well-established community (both in a temporal and spatial dimension), it may be reasonable to expect that a clear and shared notion of norm violation has shaped over time, and therefore that defectors are somewhat cautious and refrain from extracting at a very high level (such as μ_{nash}). On the other hand, a relatively new community, or one which is more spatially fragmented, may be characterized by higher defection levels (e.g. $\mu > 2$), due to the lack of clarity over what the acceptable behavior is. Figure 4 illustrates the fate of cooperation for different definitions of defection, i.e. μ levels, in order to capture these different cases in a comprehensive manner. Given the positive value of the first three terms on the right hand side of (6), (with the exception of degenerate cases), it is straightforward that the proportion of cooperators will increase ($\dot{f}_c > 0$) pro-

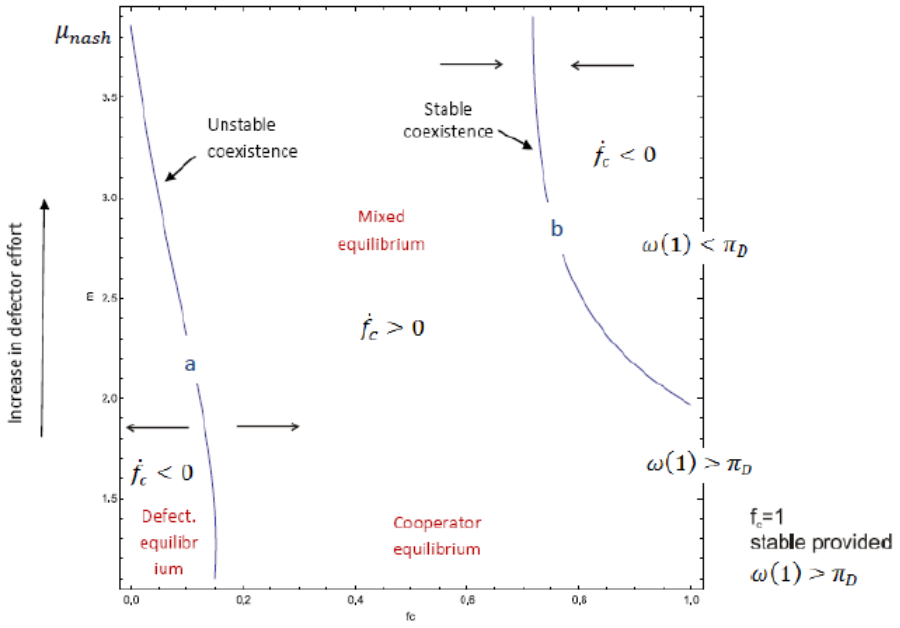


Figure 4: The $\omega(f_c^*) = \pi_d(e_d, R^*)$ loci guaranteeing coexistence of types.

vided that $\omega(fc) > \pi_d(e_d, R)$. Where the system ultimately stabilizes depends on the initial conditions: however, as we can readily observe from Figure 4, we have both areas characterized by the presence of one type of agent only and areas of coexistence. Note that with an increase in cooperators the ostracism $\omega(f_c)$ increases, but so does the defector payoff $\pi_d(e_d, R)$, because productivity increases with an increase in f_c; such interaction causes non-trivial dynamics that are analyzed here under the evolutionary lens provided by the replicator equation. In the Appendix we derive results from the stability analysis based on the linearization matrix J of the system (2)-(6), which shows that:

a) the Defector equilibrium is a stable attracting fixed point;

b) the Cooperator equilibrium is stable so long as the defector's payoff is bounded above by the full compliance ostracism costs, i.e. $\omega(1) > \pi_d$.

What happens for intermediate frequencies of cooperators? From (7) we know that, if it exists, the coexistence equilibrium must satisfy $\omega({f_c}^*) = \pi_d(e_d, R^*)$. Inspection of the curves in Figure 4 allows one to assess the qualitative features of the system resulting from the above condition: to the left of locus a, i.e. for

low initial $f_c, \omega(f_c) < \pi_d(e_d, R)$, so the system will evolve towards the stable defector equilibrium independently of μ. If, for instance, we consider defectors who extract resource according to the Nash rule ($\mu_{nash} : e_d = e_{nash}$), the equilibrium will be characterized by $\omega(0) = 0 < \pi_d(e_d, R_{nash})$ (see footnote 8). To the right of locus a, $\omega(f_c) > \pi_d(e_d, R)$, so the community of appropriators following the restrictive norm will grow larger. The system will transition towards the cooperator equilibrium when the effort difference between cooperators and defectors is not too large (low μ), as the above inequality will continue to hold until stable monomorphic cooperation obtains, with $\omega(1) > \pi_d(e_d, R_eff)$ (see Figure 4 and footnote 8). When instead effort differences are large, the proportion of cooperators will keep increasing up to a point where $\omega({f_c}^*) = \pi_d(e_d, R^*)$: At this point population composition does not change any longer and the mixed equilibrium persists. The same obtains when starting to the right of locus b; in other words, b is a stable locus of mixed equilibria. Note that this is not true for locus a, which is unstable.

3 Model extensions with variable resource replenishment rate and strategy competition

To extend the above analysis to conditions of variable resource inflow or multiple extraction strategies we developed agent-based simulations (ABM). The basic setup of the ABM closely follows the evolutionary game theoretic model, although in discrete time: individual agents are randomly matched for payoff comparison and strategy updating as in the replicator dynamics above. A time step equals the length of one replenishment cycle for the resource. Agents extract resources according to their effort strategy and produce a good that provides them with the respective payoff. At each time step two agents are matched randomly to update their utilities and strategies. When a positive payoff difference signals defection, the defector will be ostracized with a magnitude proportional to the share of cooperators in the population and the relative payoff difference between the two opponents, in equation (3). Subsequently, when his/her utility is below that of the opponent, an agent i updates his strategy by imitating the strategy of the opponent j with a probability equal to the utility difference (cf. Morgan, 2003). Letting $\Delta_i = U_i - U_j$,

$$\text{if } \Delta_i < 0 \Rightarrow e_i \rightarrow e_j \text{ with probability } \frac{U_i - U_j}{|U_i| + |U_j|} \tag{8}$$

We simulate the evolution of the population from an initial population composition and a specified effort difference between cooperators and defectors. Figure 5 shows the result of multiple simulation runs with different initial conditions and effort differences. Note that the effort difference represented on the y-axis is fixed within each simulation, while the initial proportion of cooperators represented on the x-axis is only the initial value with the final value presented in the pixels in the graph. Each square represents the average proportion of cooperators of the last 500 time steps of each run across 30 runs. With low initial proportion of cooperators and low effort difference the system converges to a Defector equilibrium. A Cooperator equilibrium emerges when initial proportions of cooperators are at least 15% and the effort difference not too large. Once effort differences increase above approximately 2 (i.e. defector effort is twice the cooperator effort) the system converges to a mixed equilibrium. When effort differences increase even more the mixed equilibrium covers the entire parameter range and the frequency of cooperators in the mixed equilibrium decreases. The simulation results are consistent with the analytical model presented in Figure 4 except for the region in the upper left corner, where in the simulations the stable defector boundary equilibrium disappears. Here, the transient dynamics characterized by stochastic updating and the discreteness of the individual agents create a situation where the resource is depleted faster than the single cooperator has a chance to update his strategy to defection. In the game theoretic model with high effort differences the unstable boundary to the mixed equilibrium approaches the stable all defector equilibrium and thus small perturbations, e.g. due to the stochastic updating, can shift the system to the mixed equilibrium.

3.1 Variable Inflow

In reality resource flows such as water flows in a landscape are rarely constant. Particularly in semi-arid regions water availability can vary drastically within and between years. Climate change is likely to increase this variability and lead to more frequent extreme events. This puts additional pressure on water users that have to cope and adapt to changing resource conditions. To assess the effect of inflow variability on the stability of cooperation we developed an ABM with variable inflow to the resource pool. The modified resource dynamics are

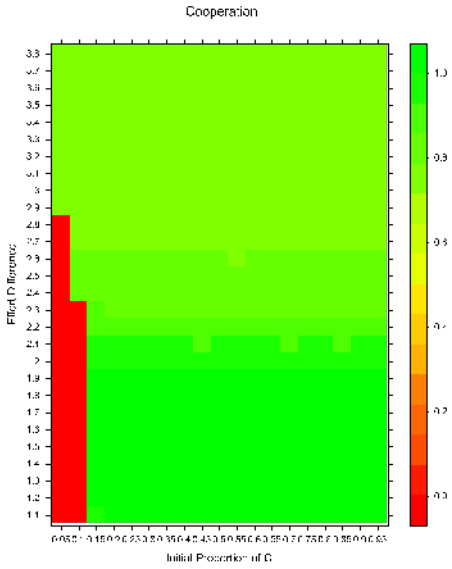


Figure 5: Frequency of cooperators for different initial f_c and different μ. Red indicates a Defector equilibrium, dark green indicares a Cooperator equilibrium. Each square represents the average of 10 different runs

given in the following relationship:

$$R_{var,t+1} = R_{var,t} + c_{var,t+1} - d_{var,t+1} * (R_t/R_{max})^2 - E_t R_t \tag{9}$$

where $c_{var,t}$ (for $t = 1, \ldots$) are independently distributed pseudorandom Gaussian values standardized to have mean c and standard deviation σ. The baseline discharge rate $d_{var,t}$ is set equal to the inflow to maintain a carrying capacity equal to R_{max}. Figure 6 shows that with an increase in variability of inflow to the common resource pool, the percentage of cooperators in the mixed equilibrium increases, thus indicating an advantage for cooperators. We explain the disadvantage for the defectors with the concavity of the resource function at $\dot{R} = 0$ (see Figure 2), which leads to a decrease in the average resource volume with inflow variability. Recalling the analysis of Section 2, letting $\tilde{R} < R$ be the perturbed resource volume at time t, we note that the corresponding payoff $\tilde{\pi}_i(e_i, \tilde{R}) < \pi_i(e_i, R)$, which implies that $\tilde{\pi}_d(e_d, \tilde{R}) < \pi_d(e_d, R)$. However, the ostracism function is independent of R, so the loci satisfying $\omega({f_c}^*) = \pi_d(e_d, \tilde{R}^*)$ will shift rightward (a and b in Figure 4). Put differently, a lower average resource volume leads to reduced payoffs for both defectors and cooperators.

Defectors, however, are also subject to ostracism which is only a function of the frequency of cooperators and thus is not affected by inflow variability. The decrease in defector payoff with constant ostracism costs decreases the frequency of defectors in the mixed equilibrium. With respect to the phase plot of the game theoretic model, the simulation results indicate that the loci of stable coexistence ($\omega(f_c) = \pi_d$) shift right and closer to the unstable all cooperators boundary equilibrium (top right corner in Figure 4).

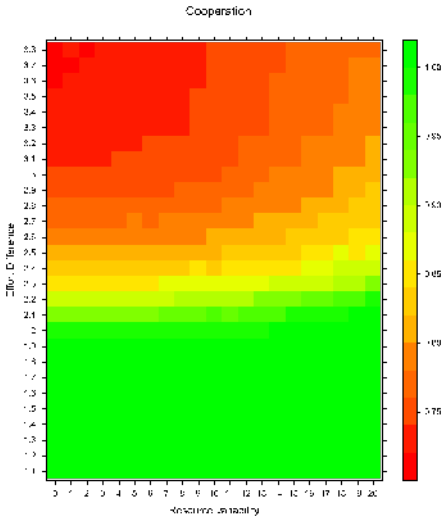


Figure 6: Frequency of cooperators with increasing variability of resource inflow and μ. The initial proportion of cooperators is set to 0.5. Here red indicates $fc = 70\%$, while green indicates 100% cooperation

3.2 Competition of Multiple Strategies

It is likely that resource users choose extraction strategies that are not confined to a set of a single cooperative and defector strategy but vary more widely. To assess how a fairness norm affects the evolution of individual extraction strategies and thus total extraction behavior of the community we developed an ABM that allows for agents with multiple strategies. Agents are initialized with a random strategy drawn from a uniform distribution on the interval $[0, \psi * E_{OA}/N], \psi \geq 1$, where E_{OA} is the aggregate extraction level that leads to open access resource exploitation. Interactions between agents are modeled as in §2, through random pairs that one at a time step update their utilities and

strategies. An agent is identified as a defector when his payoff is larger than the payoff of the opponent. Thus in this case the definition of a defector becomes dynamic and relative to the actual opponent. The magnitude of the ostracism is a function of the proportion of cooperators which is determined in the following way: cooperators are all agents who follow the norm, i.e. have an extraction strategy that is equal or lower than the socially acceptable extraction level:

$$f_c = \frac{\sum_{j=0, j\neq i}^{N} n_j}{n} \text{ for which } e_j \leq e_{eff} \tag{10}$$

The ostracism is normalized using the highest payoff available in the population at the current time step, π_{max}, instead of the (unique) defector payoff as in the game theoretic version presented in §2.

$$U_d = e_d * (\frac{f(E,R)}{E} - w) - \omega(f_c) * \frac{max\left\{\pi_d - \pi_j, 0\right\}}{\pi_{max}} \tag{11}$$

We allow for random mutations where one agent chooses a new random extraction strategy at a specified mutation rate. Figures 7 and 8 show the results of the competition of multiple strategies with different strengths of ostracism expressed through the δ parameter of the ostracism function. Each square in Figure 7 is respectively the mean or the standard deviation of 100 runs for each δ value. With increasing strength of the social norm there is a clear transition from a defector state where total effort is at open access to a Cooperator equilibrium with effort levels at or below the socially optimal level. This transition is signaled by an increase in the standard deviation of total effort which peaks around delta values of 0.5. This is a region of bistability where some simulations result in an open access total effort, while others result in lower total extraction levels (Figure 7). The average number of strategies present in the population also increases during the transition and decreases significantly with an increase in δ. Thus, with increasing strength of ostracism conformity is increased. The peak of total payoffs occurs around delta values of 1 where total effort is at the socially optimal level (0.48). With an increase in the strength of the ostracism the population converges to extraction levels that are below socially optimal indicating that the fairness norm is effective in restraining individual resource extraction but can lead to sub-optimally low resource exploitation. The latter case is reminescent of a partial tragedy of the anticommons, where too much regulation leads to too much inhibition of extractive activities, with respect to the optimum

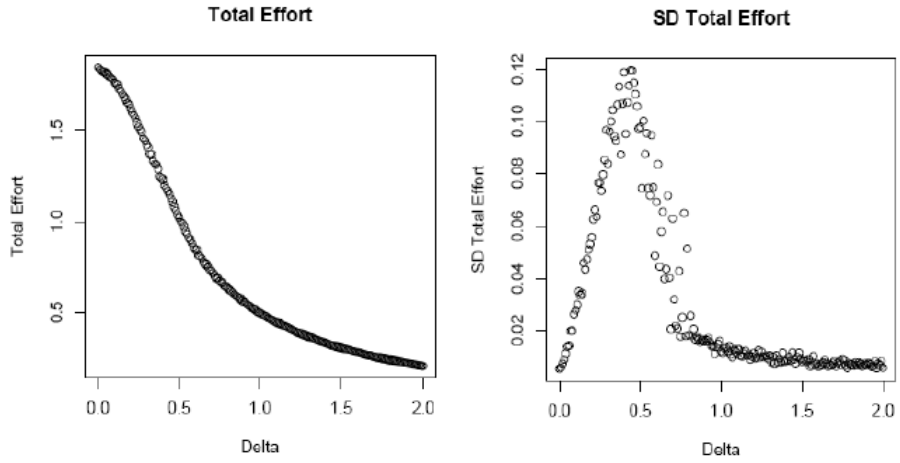


Figure 7: Average total effort (left) and standard deviation (right) of 50 runs for each δ value. Simulation initialized with 50 random uniformly distributed strategies. Mutation rate is 1 mutation every 10 time steps

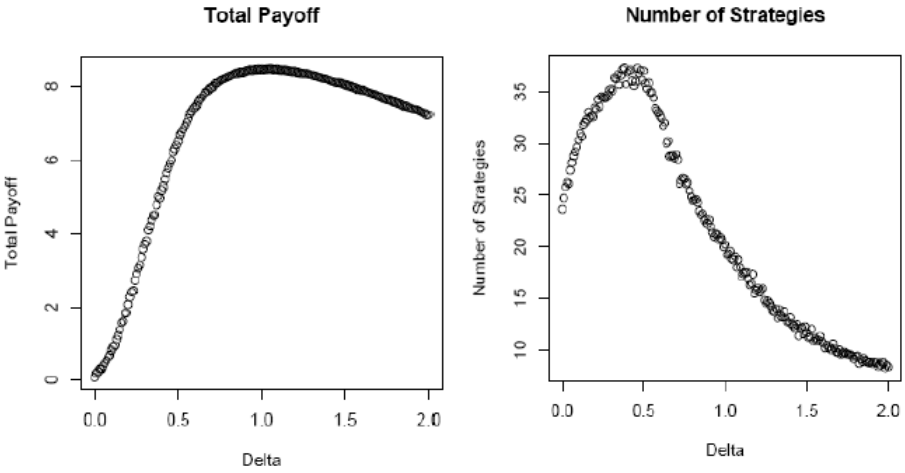


Figure 8: Average total payoff (left) and number of surviving strategies (right) of 50 runs for each δ value. Simulation parameters as in Figure 7

4 Concluding remarks

In this paper we have developed a model of community-based appropriation of a common pool resource in the presence of a norm allowing discrimination of be-

havior. Individuals departing from what the community considers as acceptable behavior (in terms of non-excessive resource extraction), are therefore subject to what we call equity-driven ostracism: a denial of support by the cooperating community which has tangible consequences on the wealth of the norm violators. Such retaliation, which may be thought of as consisting of spiteful actions (e.g. denied machinery lending and crop destruction) or social reprobation (e.g. negative gossiping and refusal to share information), depends on two factors. On the one hand the relative strength of the community of norm followers, since a larger community is assumed to be more effective at ostracizing defectors. Secondly, the intensity of the response by the community is assumed to be higher the larger the entity of the defection, which is revealed by the differences in the yield of the production, as the latter depends on the amount of resource extraction. Another noteworthy feature is that we model the coupled socio-economic and ecological dynamics of a common pool resource, such as water, that provides benefits indirectly by being utilized as an input of production rather than for its intrinsic value (as is the case for fish in fisheries). Sections 2 and 3 considered cooperation as the outcome of an evolutionary process, with successful strategies spreading in the population as a result of a process of imitation. Analytical derivations and evidence from numerous simulations allowing for complex interactions and resource dynamics lead to the following conclusions:

a) A Defector equilibrium, unregulated by the norm, is achieved only for low initial frequency of cooperators; interestingly, the basin of attraction shrinks as the gap between e_d and e_c (i.e. μ) increases, due to resource stock effects. All other initial conditions lead either (b) or (c), even when most agents initially do not abide to the norm.

b) A Cooperator equilibrium, where all abide to the norm, arises so long as the defector's payoff is bounded above by the full compliance ostracism costs: $\omega(1) > \pi_d(e_d, R_{eff})$. This is the case when the effort differences between cooperators and defectors are not too large.

c) Stable coexistence obtains when effort differences between cooperators and defectors are pronounced, sparing the latter from being eradicated: $\omega({f_c}^*) = \pi_d(e_d, R^*)$

d) Under variable resource replenishment rates, cooperators thrive better, because they can still benefit from the social capital provided by other cooperators despite a reduction in average resource volumes, while the defectors

experience a decrease in payoffs to $\tilde{\pi}_d(e_d, \tilde{R})$.

e) A competition of extraction strategies will lead to open access extraction efforts when the social norm is weak; however the system will approach optimal extraction levels rapidly when the strength of the social norm increases.

The model presented here focuses on agents harvesting from a renewable resource while facing a social norm discerning between acceptable and excessive behavior. Notwithstanding its simplicity, it allows to identify three regimes for the stationary state of the evolutionary dynamics, depending on the initial number of norm followers, the effort gap between types and the community effectiveness in enforcing the norm. In (a) the resource is severely over-harvested, a situation reminiscent of the tragedy of the commons; in (b) the resource is efficiently shared by a homogeneous population restricting use to the collectively optimal level; in (c) both type coexist and manage to partially internalize the externality. Where the system ultimately converges depends on the path followed.

Appendix

Stability analysis

In this section we shed light on the monomorphic equilibria found in §2.3 by means of the stability analysis based on the linearization matrix J of the system (2)-(6):

$$J = \begin{bmatrix} \partial \frac{\dot{f}_c}{f_c} & \partial \frac{\dot{f}_c}{R} \\ \partial \frac{\dot{R}}{f_c} & \partial \frac{\dot{R}}{R} \end{bmatrix}$$

$$= \begin{bmatrix} \varphi[\omega'(f_c)(f - f_c^2) + \omega(f_c)(1 - 2f_c) + (2f_c - 1)\pi_d] & 0 \\ nR(e_d - e_c) & \frac{-dkR^{k-1}}{R_{max}^k} - E \end{bmatrix}$$

where $\varphi = \frac{\pi_d(e_d,R) - \pi_c(e_c,R)}{\pi_d(e_d,R)}$

The bottom two entries of J are unambiguous in sign, which is respectively positive and negative. Therefore, for stability considerations, whether an equilibrium is stable depends on $\partial \dot{f}_c / f_c$. When $f_c = 0$, it reduces to $\varphi \pi_d < 0$, so $tr(J) < 0$ and $Det(J) > 0$, which means both eigenvalues of J are negative real numbers and the Defector equilibrium is a stable attracting fixed point. It can be shown that, when $f_c = 1$, $\partial \dot{f}_c / f_c = \varphi(\pi_d - \omega(1))$, which is negative

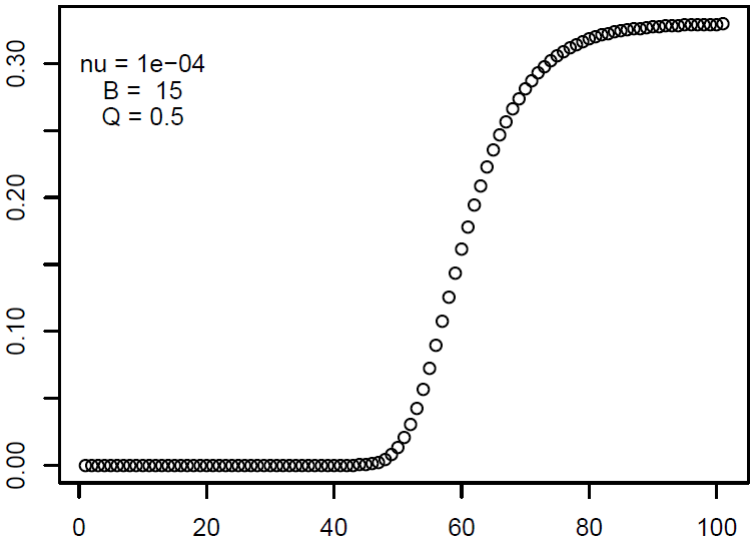


Figure 9: A threshold ostracism function

provided that $\omega(1) > \pi_d$. Thus, the Cooperator equilibrium is stable so long as the defector's payoff is bounded above by the full compliance ostracism costs.

Alternative ostracism function

The specification employed in Figure3 may be too favorable for the cooperators, since even at low f_c they are able to exert some degree of pressure on the defectors. It may be more realistic to consider the case where the community of norm followers becomes effective in ostracising the violators only beyond some threshold. For instance, when at least half of the population is comprised of cooperators. While a full account of the impact of utilizing this function in place of the one depicted in Figure 3 is not yet available, we have initial evidence suggesting that the qualitative features found in §2.3 still hold, despite a reduced parameter space supporting the Cooperator equilibrium.

References

Chen, X. et al.: Linking social norms to efficient conservation investment in payments for ecosystem services. Proceedings of the National Academy of Sciences, 106(28) (2009)

Dasgupta, P., Heal, G.: Economic theory and exhaustible resources. Cambridge University Press, Cambridge (1979)

Ibáñez, J., Martínez, S., Martínez, J.: Competitive and optimal control strategies for groundwater pumping by agricultural production units. Water Resources Research 40 (2004)

Levin, S.: Learning to live in a global commons: socioeconomic challenges for a sustainable environment. Ecological Research 21(3) (2006)

Lindbeck, A.: Incentives and Social Norms in Household Behavior. The American Economic Review 87(2) (1997)

Maier-Rigaud, F.P., Martinsson, P., Staffiero, G., Ostracism and the Provision of a Public Good Experimental Evidence. Journal of Economic Behavior and Organization (2008)

Morgan, J., Steiglitz, K.: Pairwise competition and the replicator equation. Bulletin of Mathematical Biology 65(6) (2003)

Noailly, J., Withagen, C., van den Bergh, J.:Spatial Evolution of Social Norms in a Common-Pool Resource Game. Environmental & Resource Economics 36(1) (2007)

Osès-Eraso, N., Viladrich-Grau, M.: On the sustainability of common property resources. Journal of Environmental Economics and Management 53 (2007)

Ostrom, E.: Governing the Commons: The Evolution of Institutions for Collective Action Ostrom, Elinor, Cambridge University Press (1990)

Ostrom, E.: Challenges and growth: the development of the interdisciplinary field of institutional analysis. Journal of Institutional Economics 3(03) (2007)

Sethi, R., Somanathan, E.: The Evolution of Social Norms in Common Property Resource Use. The American Economic Review 86(4) (1996)

Raakjær Nielsen, J., Mathiesen, C.: Important factors influencing rule compliance in fisheries lessons from Denmark. Marine Policy 27(5) (2003)

Taylor, P., Jonker, L.: Evolutionarily Stable Strategies and Game Dynamics. Mathematical Biosciences 40(2) (1978)

Xepapadeas, A.: Regulation and Evolution of Compliance in Common Pool Resources. The Scandinavian Journal of Economics 107(3) (2005)

Coordinating to protect the global climate: experimental evidence on the role of inequality and commitment

Alessandro Tavoni*, Astrid Dannenberg†

Abstract

Free riding is held to be the primary cause of cooperation breakdown among nonrelatives. This thwarting effect is particularly severe in the absence of effective monitoring institutions capable of sanctioning deviant behaviour. Unfortunately, solutions to global environmental dilemmas, like climate change, cannot depend on coercion mechanisms, given the transnational effects of emissions. A further complication is that, while addressing climate change requires large scale cooperation, due to the ineffectiveness of unilateral action in the face of the global nature of the problem, it yields "common but differentiated responsibilities". Such asymmetries in wealth and carbon responsibilities among the actors, and the ensuing issues of equity, might further impede cooperation. Yet, a growing literature stresses the importance of non-economic factors in explaining human behaviour; therefore, instruments that go beyond the traditional incentives might prove effective in facilitating the task. Given the empirical nature of the problem, we address it by means of controlled laboratory experiments: threshold public goods games are used to investigate the degree of cooperation achieved by groups of six participants in combating simulated dangerous climate change. While necessarily simple for the sake of tractability, these games are designed to incorporate key real-world issues, such as equity and the impact of emergent institutions based on nonbinding "pledge and review" mechanisms.

*Advanced School of Economics at the University of Venice, Cannaregio 873, 30121 Venice, Italy. Email: alessandro.tavoni@unive.it

†Centre for European Economic Research, L 7, 1 68161 Mannheim. Email: dannenberg@zew.de

1 Introduction to Chapter 3

This research aims at contributing to what remains a challenging issue across disciplines: shedding light on the drivers of cooperation among unrelated individuals in human societies. Global commons, such as the Earth's climate, represent a salient case, since the absence of property rights in valuable resources is commonly held to be a prime cause of overexploitation, potentially leading to unrecoverable resource degradation. In the absence of enforcement mechanisms, conventional game theory utilizing one-shot or repeated interactions predicts that the temptation to defect leads individuals to resource overuse, hence justifying the negative outcome generally referred to as the tragedy of the commons (Hardin, 1968). A growing body of knowledge, however, demonstrates that in many situations there is more to human behaviour than self-interest: findings from behavioural economics, evolutionary game theory and neuroscience have relentlessly made the case that human choice is a social phenomenon.[1]

This paper is concerned with the drivers of cooperation among groups of unrelated individuals faced with a coordination game requiring multilateral effort in order to reach a target and avoid losses to all members. To this end, an experiment regarding a threshold public goods game with distinctive features such as the climate change game is utilized. We have built upon the game proposed by Milinski et al. (2008) to explore some further aspects that weren't captured by the original design, and that we deem important both at the theoretical and policy level.[2] Two salient and distinguishing characteristics of the latter concern the individuals' attitude towards risk and time. On the risk–aversion side, it sets itself apart from commonly studied public goods games, as it involves investing in a public good (climate protection) in order to avoid a loss (hazardous climate change), rather than realizing a gain. Second, a relevant trait of the climate problem is the tension between avoiding incurring immediate mitigation costs by not contributing to the public good today, and the long-term preference for a sound environment. In order to incorporate salient features of the ongoing debate over how best to share the "common but differentiated responsibilities" of climate change, we focus on two aspects that are, to our knowledge, absent in the experimental literature.[3]First, we explicitly consider how it is per-

[1]A thorough review can be found in Gowdy (2008), while experimental evidence related to the present work is found in Milinski et al. (2008). For a recent empirical study showing that social norms significantly influence behaviour, effectively restraining agents from overexploiting the commons, see Chen et al. (2009).

[2]Refer to Section 2 for details about the original game and the one proposed here.

[3]Principle 7 of the Rio Declaration on Environment and Development (1992)

ceived in the presence of an asymmetric geometry for sharing the burdens of mitigation and adaptation; that is, differences in the endowments originating from contributions (or lack of thereof) in inactive rounds of play are introduced in two treatments to convey the idea of differential wealth and responsibilities to players. Second, we empower players with the ability to make nonbinding pledges before the actions are chosen. This is reminiscent of the current climate negotiations where individual nations can make pledges in an uncoordinated manner. While these announcements don't carry any commitments with them, and contrary to the rational expectations prediction, we postulate that they may facilitate the coordination among players.

The setup described below allows us to tackle important questions concerning the division of the burdens related to climate change mitigation and adaptation among developed and developing countries. Before providing the details of the game and its relation to the existing literature in Section 2, an overview of relevant features of the present debate on fair division is provided below.

1.1 Sharing the burden of climate change

In the aftermath of the 15th Conference of the Parties of the United Nations (COP 15), which took place in Copenhagen in December 2009, two issues appear to have played a determinant role in the negotiation discourse.

On the one hand, fairness considerations in burden sharing of the mitigation costs have attracted much attention in recent debates over the uneven responsibilities between developing and developed countries, the latter being historically the main contributors to climate change. An emergent key aspect that present and future climate negotiations are bound to take in account is therefore that of equity.

At the same time, given the coordination difficulties displayed by the many participants to the COP 15 (191 countries), the door has been opened to the voluntary definition of emission reduction targets by individual countries. These short-term national targets represent nonbinding pledges to reduce green-house gas (GHG) emissions (or a correlated measure such as the carbon intensity

states: "In view of the different contributions to global environmental degradation, States have common but differentiated responsibilities. The developed countries acknowledge the responsibility that they bear in the international pursuit of sustainable development in view of the pressures their societies place on the global environment and of the technologies and financial resources they command." Cf. http://www.unep.org/Documents.Multilingual/Default.asp?documentid=78&articleid=1163 to see the 27 principles.

of output) by a certain amount before 2020, with respect to a baseline year or a business as usual scenario. This approach, while not coercive, may prove successful as a first stage to achieve the global cooperation required by the global warming problem, and given the strategic nature of the interactions between sovereign countries that need to coordinate to resolve it. Below we examine these two issues in more detail, starting with equity.

Given the unprecedented rate of GHG emission growth in some newly industrializing economies, notably China, the matter of finding a fair way to share the responsibilities in the containment of global emission among countries has become central to the debate, due to the international coordination required to tackle global warming. The "wait-and-see" approach, for instance, has informed much of the United States-China exchanges on who is to be the first mover in the emission reduction game. Advocating the other country was to take the lead in terms of timing and magnitude of GHG reductions on the grounds of reciprocity considerations, the two largest emitters worldwide (each accounts for roughly one fifth of energy related global CO2 emissions) have managed to stay clear of binding commitments during the course of the George W. Bush administration. This stall has severely dented the effectiveness of the Kyoto protocol, deprived of two major actors and limited in scope to a substantially smaller emissions market.

One of the novelties of the Copenhagen Accord with respect to the Kyoto Protocol or even the Bali Action Plan from 2007, is that it more actively engages non-industrialized (non-Annex I) countries, previously confined to the role of carbon credit sellers by means of the Clean Development Mechanism. With the new document, signatories have established "Nationally appropriate mitigation actions of developing country Parties", in addition to the "Quantified economy-wide emissions targets for 2020" for Annex I countries.[4] According to these documents, at the time of writing, some 100 countries have associated themselves with the Accord, of which 75 have also issued domestic goals for mitigation actions by 2020. Interestingly, while China, India and Russia all have submitted national targets, they are the largest greenhouse gas emitters who haven't yet signed up to endorse the Accord.

From this perspective, it is clear that there is a great deal of caution in defini-

[4]These documents are contained, in Appendix II and Appendix I, respectively concerning voluntary actions by developing countries and targets for developed countries; see http://unfccc.int/home/items/5265.php and http://unfccc.int/home/items/5264.php accessed on March, 25 2010.

tion of national climate policies, even if at this stage the pledges are nonbinding; this should be no surprise, given the high level of coordination required to provide the global public good of climate protection. Inspection of the arguments given for conditional cooperation points in the direction of equity concerns of two types. On the one hand, "virtuous" players prioritizing global emissions reduction, like the E.U., try to use the leverage of differential responsibility to engage other countries in more ambitious commitments; on the other hand, developing countries with conflicting priorities, like China, appeal to "cutting emissions, providing financial support and technology transfers" from a developed country like the U.S.

We bring these two important features of climate change, namely equity and moral obligation, under the scrutiny of the experimental lens.[5] Through it, we aim to investigate whether "high emitters" will contribute more to combat climate change in a threshold public goods game where players differ in wealth (and responsibilities) and are allowed to make non-binding pledges. Section 2 provides a brief discussion of related literature along with the design of the present experiment, while Section 3 is concerned with its theoretical underpinnings, followed by a section displaying the main results. Section 5 draws some concluding remarks.

2 Experimental Design

Most experiments on public goods utilize linear public goods games, where participants have the option to invest a fraction of their endowments in a public good, by means of a voluntary contributions mechanism (see e.g. Ledyard, 1994). Typically, the returns to the investment are equally shared among the participants, according to the marginal per capita return (MPCR); for example each allocation of a Euros by each individual in the group yields a return of b>a Euros to all members of the group, i.e. MPCR is b/a Euros. We depart from this standard formulation in many ways, in order to create a setting which incorporates realistic issues faced by climate change negotiators. First, the provision of the public good is sequential, as multiple stages of contributions (10 rounds) are performed before the assessment of the group effectiveness in preventing simulated dangerous climate change. Second, the objective of the game is to avoid a loss rather than creating a surplus by contributing to a public good

[5] See Bernasconi et al. (2010) for an experimental investigation of the role of expressive obligations in public good provision.

(with higher group contributions leading to higher returns to the players). Here players' contributions to the public good make them collectively better off only insofar they are sufficient to reach a threshold (€120). All contributions below (or above) it are wasted, as they fail to secure the keeping of the private accounts by the participants (or have no additional benefit if above the threshold). This feature leads to the next salient one, concerning the probabilistic nature of the losses. To account for the uncertainty involved in climatic change, the actions of the six players forming the groups taking part in the game have consequences that are not deterministic. If they collectively fail to reach the target required to provide the climate protection public good, they will lose their savings on the private account (what is left of the initial €40 endowment after the contributions to the public good) with a probability of 50%. This level was chosen in the light of the results of the experiment by Milinski et al. (2008), which shares with us the above departures from standard public good games, and which we aim to enrich with features that will be discussed below. It is therefore worth taking a closer look at their experiment.

In a nutshell, Milinski and his co-authors implemented the above setup, with individuals deciding on each of the ten rounds of the game whether to contribute either €0 ("selfish"), €2 ("fair"), or €4 ("altruistic") to the climate account, with each group being presented with one of three different treatments corresponding to three probability of savings' loss: 90%, 50% and 10%. These yielded the following levels of success in avoiding simulated climate change: 50%, 10% and 0%. That is with the highest stakes, due to the larger gains in expected value from reaching the target, cooperation was highest and half of the participating groups where successful in collecting at least €120, while only one group out of ten succeeded in the 50% treatment and none in the one where failing groups had only a small probability of incurring the loss. Note that the last result is not surprising from a rationality standpoint, as a player selfishly contributing €0 in all rounds would have expected earnings of €36 compared to €20 and €0 by following the remaining two pure strategies of fair and altruistic contribution. Only in the 90% treatment the social optimum coincides with the fair strategy, as it would lead to certain earnings of €20 if adopted by all subjects in all rounds, compared to expected earnings of €4 if all adopt the free riding strategy and a certain outcome of €0 if they follow the altruistic strategy.

Our basic experimental design closely follows the design of Milinski et al. (2008) with six individuals playing together in a group, each endowed with €40. The players decided in each of the active rounds of the game whether

to contribute either €0, €2, or €4 to the climate account. All groups were being presented with the probability of savings' loss of 50%. After each round the players were informed about all individual contributions and the aggregate group contribution in that round as well as the cumulative past contribution of each player and the group. As in Milinski et al. (2008), players were assigned nicknames in order to keep their identity private. Since the focus of this paper is to test in the lab for the role of inequalities in informing the debate on climate change, we introduced a series of treatments aimed at capturing features of asymmetry among participants in terms of wealth, past contributions and future commitment announcements. We believe these are important facets of today's debate about sharing the burden of climate change, and designed the experiment to incorporate them in a simple manner.

In order to induce subjects to perceive the inequalities among them as the result of past actions, we modified the game described above by replacing the first three rounds with three inactive ones where half of the group had only the option of choosing a €4 contribution, while the remaining three players could only select a €0 contribution. That is, rather than externally imposing different endowments from the beginning of the experiment, players were all told they had the full €40 endowment before the start, but witnessed through the first three rounds a growing divergence between high and low contributors. As a result of these three inactive rounds, the players begin the active play consisting of seven rounds with substantial "inherited" differences: those who forcefully contributed €12 prior to round 4 had €28 left in their private accounts, while those who previously did not contribute anything to the public good found themselves with the entire endowment available for the ensuing seven rounds. We call this treatment "Base-Fair" and we expect that this setup conveys a sense of responsibility to the relatively wealthy players, as their position is due to past free-riding. This situation is reminiscent of that of global CO2 emissions, with developed countries owing much of their prosperity to past carbon-intensive industrialization, relative to developing countries with historically smaller carbon footprints and wealth.

In order to single out the effect on cooperation of the introduced asymmetry, a "Base" treatment has been performed without such unequalizing redistribution. In it, subjects go through three inactive rounds where they all have no other option than to choose the fair strategy, i.e. contributing €2 per round. While these three inactive rounds might render the fair strategy more focal, we expected to have an impact on the level of cooperation, we speculated that

our Base treatment would yield qualitatively similar results than Milinski et al. (2008), since most changes are introduced in the remaining three treatments.

Finally, we implemented two treatments in which the subjects had the opportunity to make future commitment announcements. The "Pledge" treatment introduced two pledge stages to the symmetric case while the "Pledge-Fair" treatment implemented two pledge stages in the asymmetric case. In both pledge treatments it was common knowledge that the pledges were non-binding. The first pledge stage was after the (fixed) first three rounds. The subjects simultaneously and independently announced their intended contributions for the subsequent seven rounds. Afterwards the players saw the "intended climate account" which contained the individual contributions from the first three rounds plus the individual pledges. Thereby they immediately detected whether the intended contributions would be sufficient to avoid dangerous climate change. The second pledge stage took place after round seven. Similar to the first pledge, the players simultaneously and independently announced their intended contributions for the last three rounds and were subsequently informed about the "new intended climate account" that included past contributions and the pledges. Table 1 summarizes the key features of our experimental design and the number of participants in each session.

The experiment was run in May 2010 at the MaxLab laboratory at the University of Magdeburg, Germany. In total, 240 students participated in the experiment, whereby the pool consisted of a mixture of students with an economic or business major (60%) and students with a non-economic major (40%). Most of the students were experienced as they had participated in three or more experiments before (88%) while only few students were inexperienced (12%). Sixty subjects took part in each treatment. No subject participated in more than one treatment. Sessions lasted about 60 minutes. For each session, we recruited either 12 or 18 subjects using the ORSEE software (Greiner 2004). Each subject was seated at linked computer terminals that were used to transmit all decision and payoff information. We used the Z-tree software (Fischbacher 2007) for programming. Once the individuals were seated and logged into the terminals, a set of written instructions were handed out. Experimental instructions (see the Appendix) included a numerical example and control questions in order to ensure that all subjects understood the games. At the beginning of the experiment subjects were randomly assigned to groups of six. The subjects were not aware of whom they were grouped with, but they did know that they remained within the same group of players throughout the ten rounds. After the final round, the

players were informed whether the group had successfully reached the threshold of €120. Afterwards they were asked to fill in a short questionnaire. The questionnaire was designed to elicit the players' impressions and motivation during the game, an indicator of individual risk aversion and inequality aversion as well as the general opinion about climate change policy (see appendix). At the end of the experiment, one of two table tennis balls was publicly drawn from a bag by a volunteer student. If there was the number 1 on the ball, all players in the groups that had not reached the threshold kept the money (that was left on their private account). If there was the number 2 on the ball, these players lost their money. Out of the 20 groups which did not reach the threshold 11 groups were in luck and kept their money while 9 groups were in bad luck and lost their money. No show-up fee was administered. On average, a subject earned €17.23 in the games; the maximum payoff was €40 and the minimum €0.

The money allocated to the climate account was used to buy and withdraw CO2 emission certificates traded in the European Union emission trading scheme (EU ETS).[6] If a group had successfully reached the threshold, all of the climate account money was used in this way. In case of a failing group only half of the climate account money was used for emission certificates. Thereby, we introduced a specific field context to the experiment which made the task more realistic and might increase the participants' motivation. The experimental instructions contained a short explanation of the EU ETS and the above mentioned rules (see appendix). We announced furthermore that the purchase and the suspension of certificates would be certified by a notary and that the overall amount of certificates and the notarial acknowledgment could be found on a specific website. Overall, we spent €3,255 for emission certificates which corresponds to 217 tons of CO2 given a price of 15 €/ton.[7]

Treatment	Asymmetric players	Pledge stages	Probability of climate change	No. of subjects
Base	no	no	50%	60
Pledge	no	yes	50%	60
Base-Fair	yes	no	50%	60
Pledge-Fair	yes	yes	50%	60

Table 1: Summary of experimental design

[6] For information about the EU ETS visit the European Commission official website (http://ec.europa.eu/environment/climat/emission/index_en.htm)

[7] For emission certificate prices visit http://www.eex.com/en.

3 Discussion of equilibria

As noted in Milinski et al. (2008), the multiplicity of equilibria in the game makes classification virtually impossible. The game utilized here is a modified n-person stochastic threshold public goods game, with a total of ten rounds of which only seven allow freedom of choice over the three possible actions. Given the choice of the 50% probability of loss, conditional on the group failure to collect €120, the fair strategy provides the same take home expectation than the free-rider strategy, namely €20. This implies that any average round contribution $>$ €2 is irrational, in the sense of welfare diminishing relative to not contributing anything. In fact, borrowing the wording from Milinski et al. (2008), "each course of the game that leads to exactly reaching the target sum of €120, irrespective of who[m] contributes how much as long as each player invests" at most €20, is a Nash equilibrium. Of course, depending on the round and the path that has led to it, altruistic contributions bringing the individual sum above €20 may still be optimal if successful in guaranteeing that past investments weren't wasted.

Before commenting on the impact of the three computerized rounds in §3.2, we briefly discuss the tradeoffs inherent in the game.

3.1 Game tradeoffs

For illustrative purposes, we provide an hypothetical scenario in Table 2. Assume the group has just completed round nine, with an aggregate contribution of €108 (i.e. they are on track); assume further that four players stick to €2 in round ten, unilaterally bringing the account to €116. If the two remaining players were convinced, say due to previous contribution patterns, that only the two of them would consider deviating from the fair act in the last round, they would be facing the following figures:

Ultimately, the decision depends largely, in this situation, on the degree of risk aversion and on mutual expectations. We argue that a third driver of behaviour should not be overlooked, namely moral heuristics. In particular, especially if previous departures from symmetric burden sharing introduced the need and led to altruistic acts by some of the players, inequity aversion might motivate the latter to refuse participation in an unfair outcome, even at a deer cost to them and the others. In our experimental setting, we expect these situations to arise more frequently in the treatments with initial unequalizing

	€0	€2	€4
€0	11* (116)	11* (118)	22 (120)
€2	10* (118)	20 (120)	20 (122)
€4	18 (120)	18 (122)	18 (124)

Table 2: End payoffs (and corresponding final climate account values in parentheses) to the *row player* given round-nine moves. Entries on or below the antidiagonal are certain, while the starred entries are expected values based on the 50% probability of account loss.

rounds, as they are likely to result in greater disparities among players (due to the constrained behaviour in the early rounds).[8]

Inequity aversion may be determinant in guiding the decision based on Table 2-type of scenarios. If for example a player is risk-averse but strongly resists disadvantageous inequity (has a high α parameter, in Fehr and Schmidt, 1999 terminology), he or she will be unwilling to compensate for the actions of the risk-seeker(s).

Let's return to the above example in order to evaluate how inequity aversion may steer the end result towards successful or unsuccessful coordination. In its absence, a risk-seeking player believing the opponent to be risk-averse (i.e. placing a high probability on his/her choosing €4), might be inclined to take a chance and choose €0 in the last round. Symmetrically, a risk-averse individual, say the column player, fearing to see the certainty of a gain jeopardized as a result of free-riding, may well opt for contributing €4. In that case, the two contributions would offset each other and €120 would be reached with certainty (top right entry in Table 2). This situation is reminiscent of the snow drift game, which differs from the prisoner dilemma game in that unilateral action, while not as desirable as shared cooperation, still provides a benefit to its pursuer.[9]

[8] See the discussion on group level patterns in Section 4.

[9] Kümmerli et al. (2007) argue for the omnipresence of these situations in human working life, with the following example: "two scientists accomplishing a research project would each

However, if risk aversion is dominated by inequity aversion, the column player may choose either the fair or selfish act, if believing row player to act selfishly, thus leading to the highly inefficient outcome represented by the top left and top middle cells. Highly inefficient since they don't guarantee certainty of success, notwithstanding the substantial contributions, which on average are close to €2/round per player.

3.2 Impact of the computerized rounds

As discussed in Section 2, in two treatments the players witness three rounds of unavoidable €2 contributions, while in the remaining two treatments the players undergo three unequalizing rounds resulting in half of the group being wealthier than the remaining half. At the group level, independent of the treatment, they contribute €36 to the public good before round four begins, keeping them on track with respect to the threshold. What is the impact of this mechanism on the attainable game equilibria? Let's begin by considering the case of symmetric contributions constrained to €2/round. Of the two symmetric Nash equilibria from the setup in Milinski (2008), corresponding to all players contributing €2/round or €0/round, the latter is no longer available. This difference may promote cooperation, as the unrecoverable individual contribution of €6 early in the game could in principle steer away individuals from the fair share equilibrium towards the selfish one, since they are equivalent in terms of expected payoff to the subjects.[10]

For what concerns the remaining two treatments, both symmetric Nash equilibria disappear, as not only the all selfish equilibrium is ruled out by the first three rounds (although now three players do have the option to avoid any contribution), but also the one where all players contribute two euros in each round. This since half of the group begins round four with a sunk investment of €12, while the remaining players are unbound. The difference with the previous case is stark, as it arguably introduce profound differences in the motivations of the two subgroups to provide the public good. Those who had no choice but con-

benefit if the other invests more time than oneself in the writing of the paper reporting the collaborative work. But if one of the collaborators does not contribute at all, the best option probably remains to do all the work on one's own." We believe that these tradeoffs, which also apply to the sharing of the global climate bill, are captured by the game analyzed here.

[10] In the experiment by Milinski et al. (2008), participants of the 50% treatment, which weren't bound to the fair amount in rounds one to three, contributed on average $> €1.5$/round. This suggests that the selfish Nash equilibrium was not popular even in the absence of the discussed mechanism.

tribute 30% of their endowment early on, may be more committed to going the extra step to reach the target of €20/person. The empirical question is: will the remaining players be sufficiently committed?

	Fair	Selfish
Fair	20, 20 (120)	10*, 20* (60)
Selfish	20*, 10* (60)	20*, 20* (0)

	Fair	Selfish
Fair	20, 20 (120)	10*, 17* (78)
Selfish	17*, 10* (78)	17*, 17* (36)

	Fair	Selfish
Fair	14, 26 (120)	7*, 20* (78)
Selfish	14*, 13* (78)	14*, 20* (36)

Table 3: A coordination game situation: end payoffs (and corresponding final climate account values in parentheses). Selfish refers to the strategy of giving €0 in each of the active rounds (10 rounds in the left matrix, 7 in the remaining two), Fair to giving €2/active round. While all matrices are based on an initial endowment of €40, in the games introduced here the endowment before round 4 is either €34 for all players (centre matrix), or alternatively €28 for "poor" row players and €40 for "rich" column players (right matrix). Payoffs above the antidiagonal are certain, while the starred entries are expected values based on the 50% probability of account loss.

Before turning to it, at the risk of oversimplifying the complexity of the 6-person, 10-round game, we present payoff matrices in Table 3, with the aim to highlight some key characteristics of the game in Milinski et al. (2008) and in the present work. The left matrix concerns the former, while the centre and right matrices respectively summarize the outcome of interactions in the symmetric and asymmetric games introduced here. For the sake of presentational clarity, we have simplified the analysis by assuming that two subgroups of three players choosing the same strategy form, effectively reducing the type of interactions to those present in the familiar 2x2 formulation. That is, the three players in each subgroup act identically, as if they tacitly coordinated on the same choices. Moreover, in Table 3 players can only choose between either free-riding in all rounds (Selfish strategy), or always contributing the fair amount of €2 (Fair strategy). This simplification allows analyzing the game as if it was a one shot game, where people simultaneously reason on the outcome from picking one of two strategies leading to the corresponding group level Nash equilibria (keeping in mind the above discussion on no longer attainable Nash equilibria).[11]

[11]It is important to stress again that, while the all fair-sharer equilibrium is present in all three matrices (top-left cells), the one where all players choose the selfish act in each of the ten rounds (bottom-right cell in the first matrix) is not preserved in either of the games

Comparing the three cases, we notice that, when choosing between Selfish and Fair in the respective games, best response behaviour leads to two pure strategy Nash equilibria where all players coordinate on either the selfish or the fair strategy, irrespective of which matrix we consider. However, while in the one simplifying the game in Milinski et al. (2008), both are payoff equivalent, with the Fair equilibrium being a weak Nash equilibrium and the Selfish equilibrium being strict, in the symmetric game in the centre of Table 3 the Fair equilibrium is payoff dominant (and both are strict). Lastly, in the asymmetric one, the Fair equilibrium is again payoff dominant, although it is weak, unlike the Selfish equilibrium which is strict. This analysis confirms that the games experimentally tested here can be seen as coordination games of the Stag Hunt kind, with the tradeoff betweeen social cooperation and safety being represented by the more rewarding Fair strategy vs. the safer Selfish strategy, which doesn't require cooperation to succeed.[12]

4 Results

The bird's eye view on the cooperation level across treatments is provided in Figure 1, which reports the success rate in providing the public good of climate protection. That is, for each treatment, it shows the percentage of groups who contributed at least €120 to the climate account. Inspection of Figure 1 suggests:

a) the pledges are effective tools to ease coordination among group members;

b) inequality disrupts cooperation, and more severely so in the absence of the pledges.

In the following three sections, we take a closer look at between and within treatment differences, and find supporting evidence for the above claims, as well offering explanations based on the underlying patterns.

introduced here. In other words, due to the introduction of the computerized rounds, the €0 contribution is no longer attainable in the remaining two matrices.

[12] Skyrms (2001), has the following interpretation of the game: "In the Stag Hunt, what is rational for one player to choose depends on his beliefs about what the other will choose. Both stag hunting and hare hunting are equilibria. [....] A player who chooses to hunt stag takes a risk that the other will choose not to cooperate in the Stag Hunt. A player who chooses to hunt hare runs no such risk, since his payoff does not depend on the choice of action of the other player, but he foregoes the potential payoff of a successful stag hunt. Here rational players are pulled in one direction by considerations of mutual benefit and in the other by considerations of personal risk". The game analyzed here adds a further layer of complexity, as the option that doesn't require cooperation to succeed, namely the always defect strategy labelled Selfish in Table 3, is not entirely safe due to the associated probabilistic payoff; Fair, on the other hand, is risky in terms of reliance on coordination, but safe with respect to the ensuing payoff.

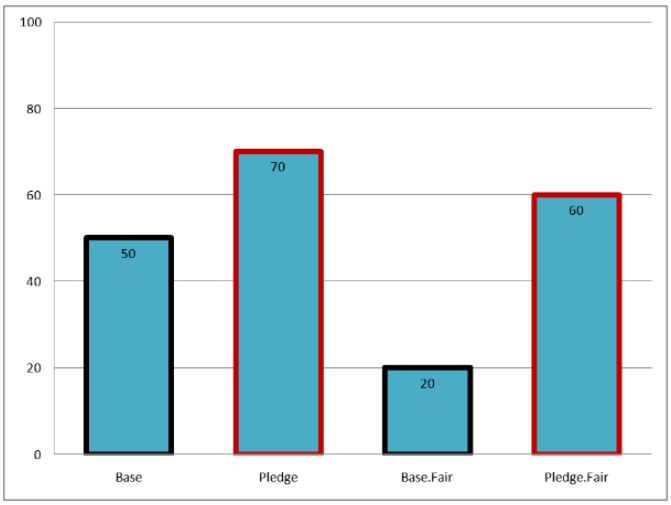


Figure 1: Success rate by treatment

4.1 Trajectories

Much of this section's analysis is based on Figure 2. In it, the contribution trajectories resulting from averaging those of the participants of the four treatments are contrasted with the symmetric trajectory that would arise if all subjects chose the fair share strategy in each round, therefore collecting €12/round. Note that each curve concerns eight rounds, the first of which represents the group contribution in round three, set by default at €36 for all treatments (see Section 2 for the experimental design), after which each subject has the freedom to choose the round contribution between €0, €2 and €4.

The experimental subjects displayed a significant amount of variation, with some groups contributing little to the public good (the group that came closest to the selfish equilibrium of €0 contribution collectively contributed only €12 in the seven active rounds), and others surpassing the threshold (the maximum was €126). Nevertheless, each treatment was characterized by substantial differences in terms of success rate in simulated climate protection. Five of the ten groups participating in Base were successful, contributing on average €122.4, while the remaining five fell short by contributing €70. The ten groups as a whole contributed €96.2 $\pm$ 32.5 (mean $\pm$ error), as shown in Figure 2. As expected, the Pledge treatment proved effective in facilitating coordination, even if based on nonbinding commitments; successful coordination on the target increases to 70%, with all groups contributing €103.6 $\pm$ 29.6, stemming from the €121.1 set

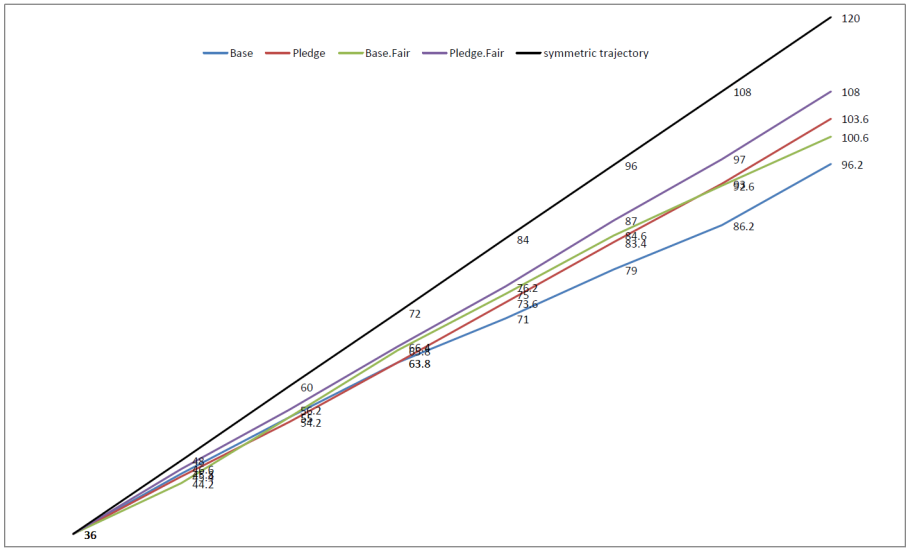


Figure 2: The contribution patterns in each treatment, starting with round 3

aside by the seven groups who reached the target and €62.7 by the remaining three.

The effect of introducing asymmetric endowments to the Base treatment is negative: compared to it, the participants of Base-Fair where 30% less successful (see Fig. 2). Interestingly, adding the possibility to make pledges again proved to be an extremely powerful tool to facilitate coordination on the threshold: Pledge-Fair groups had a success rate of 60%, which is remarkably higher than the 20% achieved by groups in the Fair treatment (and 10% higher than that of groups participating in Base, the symmetric treatment without pledges). In terms of average giving, as evident from Figure 2, participants of Base-Fair provided €100.6 ± 21.8, which is below the provision level in both pledge treatments (the highest across treatments was achieved in Pledge-Fair, with 108 ± 21.8), reflecting the positive impact of the pledges discussed above. Notably, this impact is higher when considering the asymmetric treatments (+40% success rate from Base-Fair to Pledge-Fair), with respect to the symmetric ones (+20% success rate from Base to Pledge).

What is not captured in these treatment-wise comparisons (Fig. 1 and Fig. 2) is the differences in behaviour between failing groups, which sheds light on the motivation (or lack of thereof) to provide the public good of climate protection. While in Base and Pledge failing groups provided only €70 and €62.7 respec-

tively, failing groups participating in Base-Fair and Pledge-Fair contributed a remarkable €95.5 and €88, despite the lower success rate in the latter two (-30% in Base-Fair w.r.t. Base, and -10% in Pledge-Fair w.r.t. Pledge, see Fig. 1). This evidence, together with questionnaire analysis, suggests that the role of the asymmetric endowments is twofold: it disrupts cooperation by rendering more complex coordination, but the increased failure rate is not simply the result of a decision by a larger proportion of group members to opt for a selfish strategy in the hope of high earnings. Many groups in these two treatments clearly tried to reach the €120 threshold until the last rounds, therefore increasing average contribution relative to the failing groups in Base and Pledge, who often behaved as if they tacitly agreed on gambling with the probability, due to low contributions in the early rounds. In fact 6/8 failing groups (75%) in Base and Pledge combined provided $\leq$ €70, while in the corresponding asymmetric treatments only 2/12 failing groups (17%) provided $\leq$ €70. In other words, the inequality undermined the groups' ability to combat simulated climate change, but not their motivation, which is actually higher than in symmetric treatments (cf. green vs. blue and purple vs. red lines in Fig. 2).

4.2 Contribution dynamics

Taking a closer look at Base-Fair, an analysis of the dynamics of contributions provides a perspective on the patterns behind the high number of failures that characterized this treatment. Figure 3 shows, for all treatments, the instances of selfish, fair-share and altruistic acts corresponding to giving €0, €2 and €4, respectively, in a given round. Note that, in order to have comparable figures, round four is not considered in the chart, which instead focuses on contributions in rounds five to seven and eight to ten.

The trend shaping between early and later rounds is quite pronounced: selfish acts increase on average by close to two, fair-share acts decrease by more than one and altruistic acts drop almost by one to less than one act/round in the last three rounds. This account explains the almost ubiquitous coordination failure among participants: selfish acts increase over time, while both fair-share and altruistic acts decrease over time, leaving little scope for catching up in the final rounds.

Unsurprisingly, the two treatments characterized by the highest success rate, Pledge and Pledge-Fair, owe much of it to the different dynamics, since contributions in round four where similar across all treatments (the smallest round

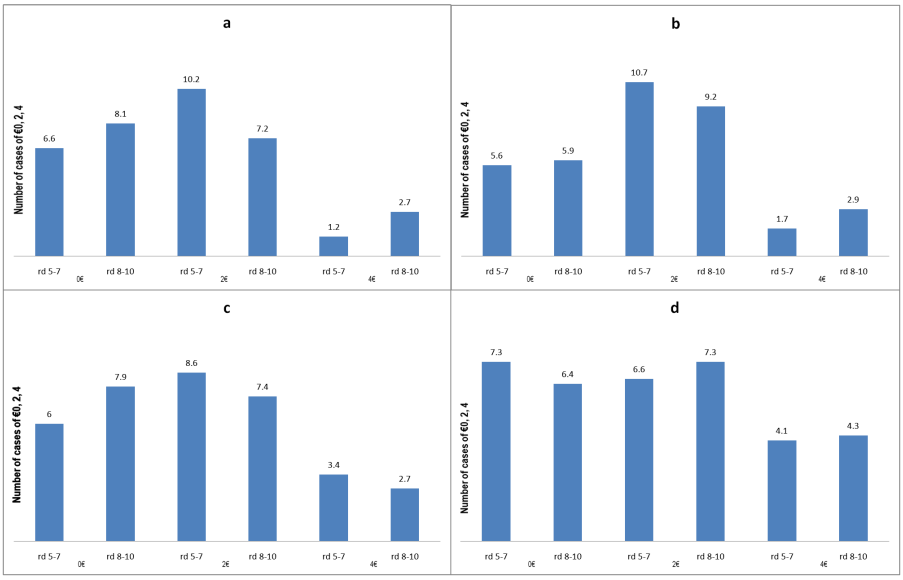


Figure 3: Contributions in early and late rounds. Amounts invested in rounds 5-7 and 8-10 to protect climate in: (a) Base; (b) Pledge; (c) Base-Fair; (d) Pledge-Fair

four contributions come from Base-Fair participants with €8.2 and the highest from Pledge-Fair participants with €10.6, relative to an overall average contribution of €9.4). Let's consider Pledge first: the 70% success rate is the result of maintaining the number of selfish acts relatively constant at a level comparable to that seen in rounds five to seven in Base-Fair (here it changes from 5.6 to 5.9), having a high (although declining from 10.7 to 9.2) number of €2 contributions, and compensating the fair-share acts decline with an increase of altruistic acts from 1.7 to 2.9 in the last three rounds.

The picture that takes shape is one with a persistent core of free-riding activity (almost two instances of €0/round in both legs), a majority of subjects contributing the fair share (more than three subjects/round in both legs) and a crucial minority of altruists giving almost €4/round in the last three rounds. Note that in the case of Pledge there is no incentive for wealth redistribution across players, as the first three rounds require players to contribute €2/round. In order to draw a comparison between the highly unsuccessful Base-Fair treatment and a more successful one, it is therefore useful to take a closer look at the dynamics in Pledge-Fair, since both are subject to three unequalizing rounds at

the beginning. Interestingly, the number of selfish acts is higher in rounds five to seven relative to Base-Fair, showing, perhaps as a result of the pledge signalling mechanism, a stronger response in the direction of equal distribution in Pledge-Fair. Put differently, as a result of the first round of pledges, those who had forcefully contributed €4 in the initial rounds may have felt more confident that the formerly free-riders would compensate prior actions by contributing €4 themselves for three rounds, and therefore engaged in more selfish acts than participants of Base-Fair. What is equally important, however, is that unlike in latter treatment, the number of selfish acts did reduce to 6.4 in the last three rounds. For what concerns the €2 count, the differences are not stark, as in the six rounds combined the Pledge-Fair participants chose the fair share close to 14 times, while the Base-Fair participants chose it 16 times. What ultimately proved to be determinant in successfully steering the Pledge-Fair cumulative contributions upwards were the altruistic acts, which in several instances sufficed to offset the selfish acts.

Compared to the close to 6 instances in the last six rounds of Base-Fair, participants of its counterpart allowing for two pledges (the last one of which before the crucial last three rounds) on average engaged more than 8 times in altruistic acts. We read this as improved coordination stemming from a commitment that, while nonbinding, nevertheless was an important vehicle of intentions among the participants. As noted before, such "lubricant of cooperation" was particularly effective in the presence of inequalities, which presumably increased the complexity of coordination by bringing fairness issues to the table, with potentially contrasting interpretations over the moral obligations stemming from them. It should be noted that the subjects took seriously the opportunity to express their planned contributions. In Pledge-Fair, for instance, the average contributions are almost identical to the corresponding pledges: between round four and round ten, contributions amounted to €72 and pledges to €71; in the last three rounds, contributions amounted to €31.8 and pledges to €32.6.

So far we have only tangentially discussed contributions in the first active round of play, namely round four. While, as noted above, variation across treatments is limited, an interesting aspect is whether there are marked differences between average round four contributions in failing groups with respect to successful ones. The answer is yes: in all treatments success in the entire game is highly linked to contributions in round four. The twenty groups that were able to coordinate to 'protect the climate' had average individual contributions of €1.9 (corresponding to €11.4 at the group level), while the remaining twenty

groups had initial individual provisions of €1.2 (corresponding to €7.3 at the group level). We therefore conjecture that the first actions carry an important weight as they signal the members' commitment in taking quantifiable efforts early on. In terms of feasible trajectories to reach the €120 target, this difference is a small burden, as it only takes slightly over one altruistic act in the ensuing six rounds to compensate the gap accumulated in round four between successful and unsuccessful groups. Yet, we argue that this lack of early initiative has deep symbolic value and explains the resulting differences in success rate. Such insight is of relevance for the current climate negotiations, and reinforces the importance of following up declarations with tangible action, especially among developed nations with higher responsibilities.

4.3 Group level patterns

Before moving on to Section 5, we will inspect behaviour in certain groups that either displayed a recurring pattern or one which is worth of notice. The first one considered in Fig. 4a belongs to the latter category. While the group, which took part in the Base-Fair treatment, got off on a good start, mimicking the symmetric trajectory in rounds four and five by providing €12 in each, and continuing to oscillate around this contribution level until round nine (where they were actually ahead by €2 with respect to the symmetric trajectory), a meagre €6 was contributed in the last round and the threshold missed by €4.

This extremely irrational behaviour, in terms of departure from payoff maximization, seems to be the consequence of an unwillingness to further invest in the climate account by those who contributed much in earlier rounds. The three players with low initial endowment due to high contributions in the first three rounds, for example, had already contributed €22 each on average by round nine, corresponding to almost €2.5/round. Their reaction was to contribute €2 collective in the last round, presumably expecting the remaining three players to provide most of the missing €10. However, the latter didn't take on the entire burden, providing only €4 collectively. This qualitative pattern took place in four of the forty groups in the sample, providing €116 or €118 by the last round. All these instances took place in treatments with endowment inequality, providing experimental evidence supporting the hypothesis advanced in Section 3, on the important role of inequity aversion in certain situations characterized by unequal burden sharing.

A somewhat diametrically opposite scenario is the one depicted in Figure 4b,

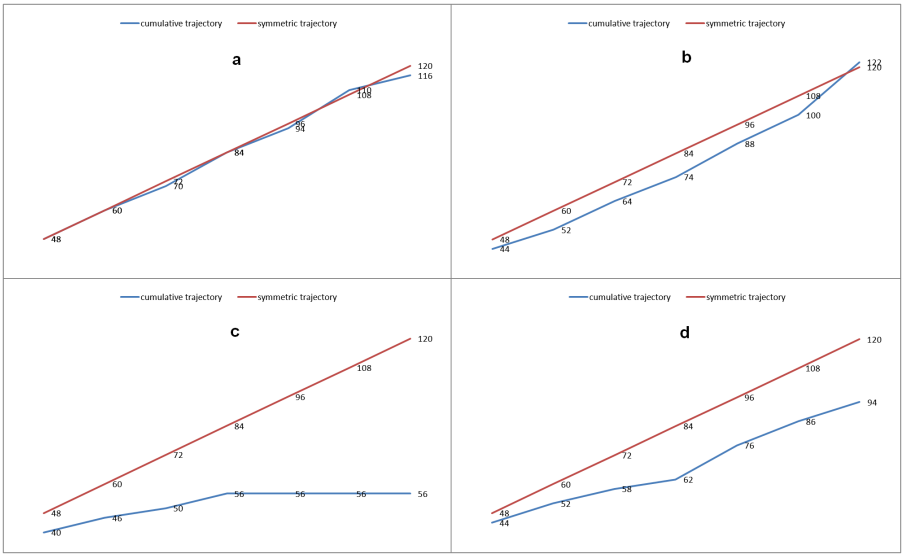


Figure 4: Selected examples of: (a) late miscoordination; (b) catching up in last round; (c) retreating early in the game; (d) still trying up to the end, but failing

where a group in the Base treatment was able to catch up, after lagging behind the symmetric trajectory in previous rounds often by €8 or even €10. Thanks to a remarkable effort in the ninth round, where all subjects contributed €4 with the exception of one who contributed €2, the group surpassed the threshold in the final round with a total contribution of €122. This qualitative pattern took place in two more groups, which successfully rebounded back from either €102 or €104 in the penultimate round.

The last two illustrations (see Fig. 4c and d), respectively taken from Pledge and Pledge-Fair, represent two failed attempts of different nature. In the first, after round seven all players abandoned hopes to provide the public good, due to the pervasiveness of selfish acts (15 in rounds four to seven), stopping at an aggregate €56. This class of group behaviour was the most frequent: 12/40 groups 'abandoned the ship' no later than in round seven, meaning that most subjects in these groups didn't make any contribution in the last three rounds, collectively providing at most €12. Notably, 67% of these instances took place in either Base or Base-Fair, indicating that the pledges promoted a sense of unity among the participants, since only 4/20 groups abandoned the ship in these treatments. The second case differs in that the non-provision of the public good does not appear to follow from an intentional decision to stop investment in

the climate account. Three players, the initially 'poor' ones, invested $\geq$ €20, while the remaining three still contributed almost €11 on average (or almost €1.1/round), which is closer to the fair share strategy than to the selfish one.

This is remarkable since, as of round nine, having the group provided only €86, it was impossible to reach the €120 target even if all had gone for the altruistic act. This suggests high motivation to protect the climate by some members, as also found in Milinski et al. (2008). In fact, in five groups at least one subject continued to submit positive contributions until the last round, even if the target was beyond reach. Again, these groups weren't evenly distributed across treatments: 80% of them took part in one of the two pledge treatments, suggesting a positive effect of the pledging ability on motivation to protect the climate.

In addition to the cases discussed above, and depicted in Figure 4, a last one deserves attention, due to its frequent appearance (9/40 groups) and theoretical relevance, the group level Nash equilibrium. By it we mean that the group as a whole successfully coordinated on a provision of precisely €120, whether or not the burden was symmetrical shared among the members.[13] In fact, only in two instances did each member contribute €20 overall (one in Base and one in Pledge), and in one of these they all played the symmetric fair-sharer strategy of always contributing €2. Braking down these instances by treatment sheds further light on the positive role of the pledges as a coordination mechanism: 6/9 Nash equilibria were achieved in pledge treatments.

5 Conclusions

In this paper we have experimentally explored the relevance of equity and commitment issues in affecting the subjects' willingness to contribute to a public good framed in terms of avoidance of dangerous climate change. The main purpose of the project was to address the question: *Will the most responsible actors contribute more to combat climate change in a public goods game experiment where players differ in wealth (and responsibilities) and are allowed to make non-binding pledges?*

Given the lack of scientific consensus on who should bear the burden of adaptation and mitigation costs, providing empirical evidence on the driving forces behind cooperation in a setting designed to mimic inequalities and bargaining

[13] This is a loose interpretation of the definition of Nash equilibrium given in Section 3, which requires symmetric behavior.

possibilities faced by actors involved with climate change, should be fruitful also from a policy perspective.

GHG emissions' very characteristic of affecting the welfare of individuals unequally, due to the uneven distribution of climate change impacts, make it a textbook example of transnational externality. Additionally, the asymmetric geometry of global emissions introduces the possibility to argue in essentially opposite directions on the grounds of fairness motives. Developing countries may insist on the importance of past emissions to justify their unwillingness to take action, while developed countries can appeal to the relevance of current emissions, generally higher in transitioning economies, to refute to take lead in mitigation actions. Such asymmetries may lead to "political lock-ins" that are detrimental to the establishment of a global agreement to curb emissions (Halsnæs et al., 2007). The game introduced here allows capturing relevant aspects concerning both the tension between collective good and free riding on the efforts of others (e.g. benefitting from polluting activities without internalizing the externality), as well as the potentially disruptive role of uneven wealth and responsibilities arising from past activities. Moreover, by introducing the faculty to make nonbinding pledges on future contributions to the public good, we have tested for the role of this institution in promoting cooperation and mitigating the problems arising from the above mentioned inequalities.

The empirical answer to the question stated above is generally no: initially wealthier subjects were often unwilling to compensate for past, "inherited", actions which had benefited them at the expense of the common good. Such resistance, much to the frustration of the remaining subjects who expected initiative on the part of the wealthy, accounted for the frequent coordination failures in the asymmetric treatments. In all twelve instances (out of twenty participating groups) where the target sum was not provided, there was an unfavourable contribution imbalance for those who had been bound to the altruistic act in the first three rounds, who ended up on average paying 60% of the bill. Not surprisingly, the burden was shared evenly in the remaining eight successful groups, with both subgroups contributing 50% of the sum.

More generally, this game captures trade-offs that are particularly salient for the issue of climate change mitigation. For instance, carbon leakage, broadly conceived as the potentially benefit-offsetting externality of climate policy in one country, in the event that its carbon dioxide emission reductions trigger an increase in another country's emissions, is a serious threat to international cooperation on global emission reduction. The ensuing trade-off between common

good and self-interest can be seen as a coordination problem where cooperation by one player favours defection by another.[14]

To the end of analysing such tensions, the climate game empirically tested here is a promising tool; notwithstanding its simplicity, it provides insights into many aspects that are crucial to climate change and cooperation at large. We have built upon the game proposed by Milinski et al. (2008) to explore some further aspects that weren't captured by the original design, and that we deem important both at the theoretical and policy level. In particular we have focused on: (i) introducing asymmetries among players by means of a novel unequalizing mechanism in the first three rounds; (ii) allowing players to make pledges concerning future contributions.

While neither feature alters the game structure in terms of the group trajectory required to reach the threshold for climate protection, they both have a significant impact on the groups' success rate. Asymmetries undermined coordination, especially in the treatment where subjects had no signalling mechanism beyond contributions, in which 80% of the groups failed to reach the target sum. Pledges, on the other hand, proved to be an effective lubricant of cooperation, halving the percentage of failures in the treatment with endowment inequalities. Both in the baseline and across all treatments, the rate of success was 50%, a remarkably high level considering the instability of the fair share Nash equilibrium and the previous findings of 10% cooperation by Milinski et al. (2008). With respect to the latter, the higher cooperation may stem from design and subject pool differences (see Section 2 and the Appendix for details on the design and for the complete instructions translated from German). As for the former, we argue, in accordance to much of the experimental literature, that human behaviour is guided by a rich set of heuristics that may interfere with expected payoff computations, steering decisions away from the rationality prescriptions. We have discussed two such heuristics, risk aversion and inequity aversion; data and questionnaire analysis suggest that both play an important

[14] On the ambiguous relationship between carbon leakage and environmental coalition stability, Botteon and Carraro (1997) argue the following. "On the one hand, carbon leakage tends to reduce the size of stable coalitions and even the likelihood of observing a stable coalition at the equilibrium because it reduces the coalition benefit and increases the incentive to free ride when the coalition is small. On the other hand, carbon leakage increases the return from large coalitions, and decreases the return from free-riding when the coalition is large. Therefore, carbon leakage, if sufficiently large, can induce the formation of large environmental coalitions. Hence, there may be two equilibrium coalition structures: one formed by a small coalition (or by the non-cooperative equilibrium) and one formed by the grand coalition (or a very large one). How to move from one equilibrium to the other is a matter of coordination, which demands for new international institutions."

role in explaining the observed departures from best-response behaviour.

Appendix

Experimental instructions for the treatments Base-Pledge and Pledge-Fair

Welcome to the experiment!

1. General Notice

In this experiment you can earn money. To make this experiment a success, please do not talk to the other participants at all or draw any other attention to you. Please read the following rules of the experiment attentively. Should you have any questions please signal us. At the end of the instructions you will find several control questions. Please answer all questions and signal us when you have finished. We will then come to you and check your answers.

2. Climate Change

Now we will introduce you to a game simulating climate change. Global climate change is seen as a serious environmental problem faced by mankind. The great majority of climate scientists expect the global average temperature to rise by 1.1 to 6.4 degrees Celsius until the year 2100. There is hardly any denial that mankind largely contributes to climate change by emitting greenhouse gases, especially carbon dioxide (CO2). CO2 originates from burning of fossil fuels like coal, oil or natural gas in industrial processes and energy production, or combustion engines of cars and lorries. CO2 is a global pollutant, i.e. each quantity unit of CO2 emitted has the same effect on the climate regardless of the location where the emission has occurred.

3. Rules of Play

In total, 6 players are involved in the game, so besides you there are 5 other players. Every player faces the same decision making problem. At the beginning of the experiment you will receive a starting capital (= EUR 40) credited to your private account. During the experiment you can use money from your account or not. In the end your account balance will be paid out to you in cash. You will be making your decisions anonymously. To guarantee for this you will be assigned a nickname for the playing time. The nicknames are the moons of our solar system (Ananke, Telesto, Despina, Japetus, Kallisto or Metis). You will find your name on the lower left side of your screen. During the course of the experiment you will be playing exactly 10 climate rounds. In

these rounds you can invest into the attempt to protect the climate and to evade dangerous climate change. Among others, dangerous climate change will result in significant economic losses which will be simulated in this experiment. In each climate round of the game all six players will be asked simultaneously:

"*How much do you want to invest into climate protection?*"

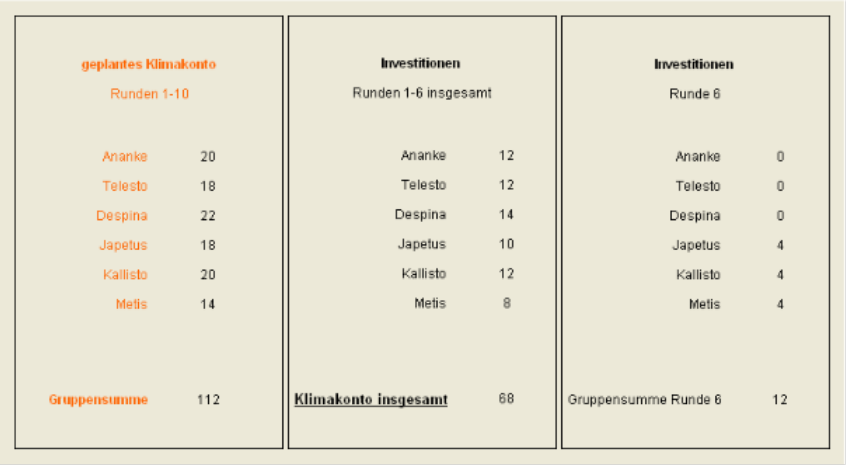
Possible answers are EUR 0, 2 or 4. Only when each player has made his choice, all decisions will be displayed simultaneously. After that the computer will credit all invested amounts to an account for climate protection ("climate account"). At the end of the game (after exactly 10 rounds) the computer will compare the climate account balance with a predetermined amount (= EUR 120). This amount must be earned to evade dangerous climate change. It will be earned if every player averagely pays EUR 2 per round into climate protection. If this is the case, EUR 12 are be paid into the climate account per round. If the necessary EUR 120 have been earned, all players will be paid out the amount remaining on their private accounts. The remaining amount consists of the starting capital of EUR 40 minus the sum paid into the climate account. If the necessary EUR 120 have not been earned, dangerous climate change will occur with a probability of 50% (in 5 out of 10 cases) and this will result in significant economic losses. If this probability arises you will lose all money left on your account and no one will be paid out anything. With another probability of 50% (in 5 out of 10 cases) you will keep your money and will be paid out the amount on your private account after the game. We will draw the probability by lot in your presence. The payout will be made anonymously. Your fellow players will not learn about your identity. Please note the following two particularities in the game: First, the decisions of the six players in the first three rounds are predetermined by the computer. Meaning, you - and your fellow players - cannot decide freely how much you want to invest into climate protection in the first three rounds. You will be offered an option instead which you have to choose.

Please note that the predetermined investments of the first three rounds will already change the amounts on the climate account and the players' accounts! Starting in round 4 you will decide freely which amounts you want to invest into climate protection. Second, all players can issue declarations of intent about how much they want to invest into climate protection in the following rounds. The declarations are not binding for the investment decisions in the following rounds. The first declaration of intent is issued after round 3. All players will simultaneously state how much they plan to invest into climate

protection in the next seven rounds in total. When all players have stated their declarations of intent, the "planned climate account" will be displayed. The planned climate account shows the investments of each player of the first 3 rounds plus the investments planned for the remaining seven rounds. After round 7 all players will be given the opportunity to revise their declarations of intent. All players then simultaneously state their planned total investments into climate protection for the next three rounds. When all players have stated their declarations of intent the "newly planned climate account" will be displayed. The newly planned climate account shows how much each player has already invested in the first seven rounds plus the planned investments for the remaining three rounds.

4. Example

In this example you see the decisions made by the six players in one round (round 6).



geplantes Klimakonto Runden 1-10		Investitionen Runden 1-6 insgesamt		Investitionen Runde 6	
Ananke	20	Ananke	12	Ananke	0
Telesto	18	Telesto	12	Telesto	0
Despina	22	Despina	14	Despina	0
Japetus	18	Japetus	10	Japetus	4
Kallisto	20	Kallisto	12	Kallisto	4
Metis	14	Metis	8	Metis	4
Gruppensumme	112	Klimakonto insgesamt	68	Gruppensumme Runde 6	12

The column on the right side ("Investitionen Runde 6") shows the investments made in the current round. Players Ananke, Telesto and Despina have not paid anything into the climate account, whereas players Japetus, Kallisto and Metis each have paid EUR 4. In total EUR 12 have been paid and by that been credited to the climate account. The column in the middle ("Investitionen Runden 1-6 insgesamt") shows the total investments made by each player in rounds 1-6. Players Ananke, Telesto and Kallisto each have paid EUR 12 into the climate account in the first 6 rounds. Despina has paid EUR 14, Japetus EUR 10 and Metis EUR 8 in the first six rounds. By that a total of EUR 68 has been paid into the climate account.

The column on the left ("geplantes Klimakonto Runden 1-10") shows the planned climate account after the first declaration of intent. The value stated per player shows the investments made in the first three rounds plus the planned investments for the remaining seven rounds. Exactly this information will be displayed after each climate round.

5. Usage of the Money on the Climate Account

If the necessary EUR 120 have been earned to evade climate change, we will buy CO2 emission certificates of the total amount on the climate account and retire them. If the necessary EUR 120 have not been earned, we will use half of the amount on the climate account to buy CO2 emission certificates and retire them (we will keep the rest of the money). By purchasing and retiring the CO2 emission certificates we contribute to the abatement of climate change. We will now explain you how this works: In 2005 the European Union has implemented the emissions trading system for carbon dioxide (CO2). Emissions trading is the central instrument of climate policy in Europe. It follows a simple principle: The European Commission, together with the member states, has determined the amount of CO2 to be emitted altogether in the respective sectors (energy production and energy intensive industries) until 2020. This total amount will be distributed to the companies by the state in the form of emission rights ("certificates"). For each quantity unit of CO2 emitted, the company has to give a certificate to the state. The certificates can be traded between companies.

For each quantity unit of CO2 emitted e.g. by a power plant, the plant operator has to prove his permission to do so in the form of a certificate. This leads to an important consequence: If the total amount of certificates is reduced, the total emissions will be lower, simply because plant operators do not possess enough emission allowances. That means if a certificate for one quantity unit is obtained from the market and is being "retired" (i.e. deleted) the total CO2emissions are reduced by exactly this quantity amount. The opportunity to retire certificates actually exists in the framework of the EU Emissions Trading System. In Germany the German Emissions Trading Authority (DEHSt) regulates Emissions trading. The authority holds a retirement account with the account number DE-230-17-1. If certificates are transferred to this account they will be withdrawn from circulation, i.e. deleted, by the end of each year. ZEW has opened an own account at the DEHSt (DE-121-2810-0). The purchasing and retiring of the certificates will furthermore be attested by a notary public. Summarizing: if all players have for example paid a total of EUR 120 into the climate account, we will buy certificates for about 8 tons of CO2 (the price per

ton is currently at about EUR 15). This equals the emissions of a ride in a VW Golf (1.4 TSI) one and a half times around the world.

6. Control questions

If you have finished reading the instructions and do not have questions, please answer the following control questions.

a. Which total amount does each player have to averagely invest into climate protection in the 10 rounds to evade dangerous climate change (please tick the according box)? O EUR 12 O EUR 20 O EUR 40 O EUR 120

b. Please assume that the necessary amount of EUR 120 to evade climate change has been earned and you have invested a total of EUR 16 in the 10 rounds. How much money will you be paid out? My payout is EUR _____.

c. In how many rounds can the players decide freely about their investments into climate protection (please tick the according box)? O in 3 rounds O in 5 rounds O in 7 rounds O in 10 rounds

d. Please refer to the example stated under point 4 for the numbers. What do the balances on Despina's and Metis' private accounts state? Despina's balance states EUR _____. Metis' balances states EUR _____.

e. Please refer to the example under point 4 again. How much would the group have to pay into the climate account in the next four rounds in total to abate dangerous climate change (please tick the according box)? O EUR 12 O EUR 52 O EUR 68 O EUR 120

f. When do the players state their first declaration of intent and when can they revise this declaration? First declaration after round: _____. Revision after round: _____.

g. In your first declaration of intent after round 3 you are asked to state how much you want to invest in climate protection in the following seven rounds in total. If you want to invest averagely EUR 2 per round, which amount would you have to state in your declaration of intent (please tick the according box)? O EUR 2 O EUR 12 O EUR 14 O EUR 20

h. Are the declarations of intent binding for the investment decisions in the following rounds (please tick the according box)? O Yes O No

i. Please refer to the example under point 4 again. What do the figures in the left column "Planned climate account" stand for (please tick the according box)? O the invested amounts of the first three rounds O the planned investments for the last seven rounds O the invested amounts of the first three rounds plus the planned investments for the last seven rounds

j. Please refer to the example stated under point 3 for the numbers again. Please assume that all players adhere to their declaration of intent (see "geplantes Klimakonto"). Would the investments be enough to evade dangerous climate change (please tick the according box)? O Yes O No

k. Please assume that the necessary amount of EUR 120 has not been earned. With which probability will you lose the remaining amount on your private account (please tick the according box)? O 10% O 30% O 50% O 70% O 90% O 100%

If you have answered all control questions, please signal us. We will come to you and check the answers. After having checked the answers of all players and there are no remaining questions, the game starts. Good Luck!

Questionnaire

Question		Answer	No.	%
(1) Do you agree with the following statement? "Those who began in round 4 with a starting capital of EUR 40 should pay more into the climate account in the following seven rounds than the other players."		Agree	91	75.83
		Disagree	12	10.00
		Neither	17	14.17
(2) Please assume that three players of a group begin in round 4 with a starting capital of EUR 40 (because they have not paid anything into the climate account yet) whereas the other three players begin with a starting capital of EUR 28 (because they have paid EUR 4 into the climate account in each of the first three rounds).	What would you consider a fair average investment for the following seven rounds for those beginning with EUR 40?	0	2	0.83
		1	2	0.83
		2	30	12.50
		3	190	79.17
		4	16	6.67
	What would you consider a fair investment for the following seven rounds for those beginning with EUR 28?	0	9	3.75
		1	143	59.58
		2	85	35.42
		3	3	1.25
		4	0	0.00
(3) Please try to remember the decisions made by your fellow players during the game. In your opinion, which players have been motivated by following reasons? Please write one or more names next to each motive. Do you think there were any other motives for your fellow players besides the given? Possible motives are - Monetary self-interest - Fairness consideration - Advancement of the common coordination process - Other motives (please specify and state name)				
(4) Please briefly describe the three most important reasons for your investment decisions in a descending order of importance. Possible examples are: - Group or own investments in the *preliminary round,* - Cumulated group or own investments starting in *round 4,* - Cumulated group or own investments starting in *round 1,* - Monetary self-interest, - Fairness consideration, - Achievement of the EUR 120 limit, - Adherence to declarations of intent, - Other reasons (please state).				
(5) What has been your motivation for your investment decision in the last round (round 10)? Please state your three most important reasons in a descending order of importance (for possible answers see previous question)				
(6) If you were to play the game again, would you make different decisions? Please state your three most significant changes in a descending order of importance.				
		∑	240	100.00

Notes: Question 1 was asked in the asymmetric treatments Base-Fair and Pledge-Fair only. Question 2 was asked in all treatments; therefore it was hypothetical in the symmetric treatments Base and Pledge while it was real in the asymmetric treatments. No responses are provided for the open questions 3-6.

Table 4: Questionnaire and responses – Part I

Question	Answer	No.	%
(7) Please imagine the following situation: You have EUR 40. With a probability of 50 % you will lose all EUR 40. You could abide the risk by giving away EUR 20 of the EUR 40. Would you pay EUR 20 to avoid the risk?	Yes	165	68.75
	No	22	9.17
	Indifferent	53	22.08
(8) Did you ever donate money of goods to a charity organisation?	Often	14	5.83
	Sometimes	77	32.08
	Rarely	102	42.50
	Never	47	19.58
(9) Do you agree with this statement? "I think social differences should be levelled out more in Germany."	Agree	110	45.83
	Disagree	47	19.58
	Neither	83	34.58
(10) Do you think the problem of global climate change is being estimated correctly or not? In my opinion, the Problem is being	Rather overestimated	51	21.25
	Rather correctly estimated	83	34.58
	Rather underestimated	89	37.08
	I don't know	17	7.08
(11) In your opinion which challenges in Germany are currently the greatest? Please state the three greatest challenges in a descending order of importance.	Old age provisions	18	7.50
	Unemployment	48	20.00
	Poverty	6	2.50
	Educational policy	66	27.50
	Energy supply	3	1.25
	Health care	3	1.25
	Climate protection	13	5.42
	Crime	1	0.42
	Social security	4	1.67
	Fiscal policy	6	2.50
	Terrorism	0	0.00
	Environmental protection	3	1.25
	Economic upturn	40	16.67
	Immigration/Integration	7	2.92
	Other (please state below)	22	9.17
(12) Which of the following guiding principles describes your understanding of fairness best in the context of international climate negotiations?	Countries with high emissions in the past should reduce more emissions.	56	23.33
	Countries with high economic performance should reduce more emissions.	53	22.08
	Countries should reduce their emissions in such a way that emissions per capita are the same for all countries.	41	17.08
	Countries should reduce their emissions in such a way that the emissions percentage is the same for all countries.	53	22.08
	Other principle (please specify)	37	15.37
(13) What are the reasons for your answer in the previous question? Please state the three most important reasons in a descending order of importance.			
	∑	240	100.00

Notes: The responses to question 11 refer to the first of the three greatest challenges. No responses are provided for the open question 13.

Table 5: Questionnaire and responses – Part II

Questionnaire analysis

After the experiment subjects were asked to fill in a questionnaire about the motivation for their contribution decisions during the game and their general opinion about climate change (see appendix). Overall, the subjects appear to take climate change seriously. About 5% of the subjects think that climate protection is currently the greatest challenge in Germany. Out of 15 possible challenges for the German policy climate protection ranks sixth. However, the magnitude of the problem is seen very differently: about 21% think that the problem of global climate change is being rather overestimated, 35% think that it is being correctly estimated, 37% think that it is being underestimated, and 7% do not know. The subjects also differently evaluate the equity principles that may guide international climate agreements: 23% support the polluter-pays principle, 22% support the ability-to-pay principle, 17% favor the egalitarian principle, 22% prefer the sovereignity principle, and 16% support another principle.

The summary statistics of the players' motivation for their contribution decisions during the game are more complicated because on the one hand we used open questions to elicit the motives and on the other hand the motives obviously depend on the respective group performance. The qualitative categorization of responses reveals that the majority of players is primarily motivated by the achievement of the threshold (43%), fairness considerations (18%), material self-interest (15%), and the past group performance (14%). Understandably, the poor players in the asymmetric treatments Base-Fair and Pledge-Fair care more about fairness than the rich players (22% versus 15%) and more about the past group performance (27% versus 14%). About 6% of all subjects state that they are particularly motivated by the climate protection realized through the purchase and retirement of the CO2 certificates. In the final round the players are primarily motivated by the achievement of the threshold (42%), material self-interest (18%), the hopelessness to reach the threshold (14%), and fairness considerations (11%). The self-reported motives are in line with the actual behavior in the game, e.g. people stating that fairness was the most important reason often contributed €20 to the climate account while people stating the self-interest was their primary motive mostly gave less than €20. The self-reported motives furthermore help to understand why some groups did not reach the threshold. Comparing the successful groups that reached the threshold and the groups that did not, fairness considerations were more important for the suc-

cessful groups (23% versus 13%) as well as the achievement of the target (52% versus 35%) while self-interest (9% versus 20%) and the past group performance (8% versus 21%) were less important.

In order to elicit players' fairness perceptions, the subjects in the asymmetric treatments were asked whether they agree with the following statement: "Those who began in round 4 with a starting capital of €40 should pay more into the climate account in the following seven rounds than the other players". Overall, 76% of subjects agree with that statement, 10% disagree, and 14% neither agree nor disagree. However, there are significant differences between poor and rich subjects: out of the poor players, 90% agree, 5% disagree and 5% do neither of them while out of the rich players only 62% agree, 15% disagree and 23% do neither of them. In another question, subjects were asked "What would you consider a fair average investment for the last seven (active) rounds for those beginning with €40 and for those beginning with €28?" Possible answers include €0, €1, €2, €3, and €4. Almost all of the poor players (95%) perceive €3 as the fair amount for the rich players while only 72% of the rich players share this perception. Similarly, only 23% of the poor players perceive €2 as the fair average contribution for the poor players while 42% of the rich players state that this would be the fair amount. These specific amounts (€3 for the rich and €2 for the poor) are particularly important because they reflect the application of the different equity principles. In our game, the egalitarian rule, the polluter-pays rule and the ability-to-pay rule are equivalent: according to these principles the rich (and responsible) players should compensate for the inactive rounds where they gained their wealth without contributing to climate protection. In order to equalize the players' contributions and payments the rich should contribute €20 in the active rounds, i.e. on average €3 per round. As opposed, the sovereignity rule does not consider the players' wealth or responsibility but rather requires the same contribution during the active rounds, i.e. €2 per round for the rich as well as for the poor players. In fact, a couple of rich subjects argued that the assignment of roles was just bad luck or good luck and that the €2 contribution per round and player was a fair burden sharing. Hence, our game as much as the real climate negotiations allow for different notions of fairness. The players tend to pick the notion that is in their best interest ("self-serving bias") meaning that the implementation of that notion would generate least costs for them. This self-serving bias in the perception of fairness has been also observed in the real climate negotiations (Lange et al. 2010) and it obviously deteriorates the chances for effective coordination and cooperation.

References

Bernasconi, M. et al., 2010. "Expressive" Obligations in Public Good Games: Crowding-in and Crowding-out Effects. Working Paper

Botteon, M., Carraro, C. (1997), Burden-Sharing and Coalition Stability in Environmental Negotiations with Asymmetric Countries, in Carraro, C. ed., International Environmental Agreements: Strategic Policy Issues, E. Elgar, Cheltenham.

Buckley, E., Croson, R.,2006. Income and wealth heterogeneity in the voluntary provision of linear public goods. Journal of Public Econ

Chan, K. et al., 1996. The voluntary provision of public goods under varying income distributions. Canadian Journal of Economics

Croson, R., Marks, M., 2001. The effect of recommended contributions in the voluntary provision of public goods. Economic Inquiry 39(2): 238-249.

Fehr, E., Schmidt, K., 1999. A Theory Of Fairness, Competition, And Cooperation. The Quarterly Journal of Economics 114(3): 817-868.

Hardin, G., 1968. The Tragedy of the Commons. Science 162 (3859), 1243

Halsnæs, K., Olhoff, A., 2005. International markets for greenhouse gas emission reduction policies–possibilities for integrating developing countries. Energy Policy 33(18)

Ledyard, J. 1994. Public Goods: A Survey of Experimental Research, Econ-WPA.

Maurice, J. et al., 2010. Income Redistribution and Public Good Provision: an Experiment. Working paper

Milinski, M. et al., 2008. The collective-risk social dilemma and the prevention of simulated dangerous climate change. Proceedings of the National Academy of Sciences, 105(7), 2291-2294.

Skyrms, B., 2001. The Stag Hunt. Presidential Address of the Pacific Division of the American Philosophical Association, in Proceedings and Addresses of the APA. 75: 31-41.

Conclusions to the dissertation

The three chapters presented here encompass issues of decision making under various challenges which are prone to cause an inherent tension between self- and other-regarding behaviour. Each essay sets to explore its impact on human interactions by resorting to different game theoretic techniques.

Chapter 1 introduces inequity aversion considerations for equilibrium concepts such as impulse balance equilibrium and quantal response equilibrium. The success of the concepts to replicate experimental data, and their predictive powers are tested on 2×2 games with a unique equilibrium which is in mixed strategies. A ranking of the concepts is produced, with equity-driven quantal response equilibrium (EQRE) being the best.

Chapter 2 is concerned with a model of evolution of norm compliance in the commons, which is analysed by combine insights and techniques from two strands of literature, commons management and evolution of cooperation. The main results are that Equity-driven ostracism can promote full cooperation provided that the violation of the norm is not excessive and that enough members of the community engage in ostracism. In addition, for larger violations coexistence emerges among the two populations, with sharp transitions between non-cooperative and cooperative outcomes occurring when the stringency of social norm increases. Lastly, by means of agent based simulations, realistic features are injected in the model, such as resource uncertainty and competition among multiple extraction strategies.

Lastly, an experiment concerning a threshold public goods game framed in terms of a climate change control is presented in Chapter 3. Its main contribution is to extend the existing literature, by introducing an original mechanism which results in differing endowments and responsibilities among the players. This is achieved by denying players the freedom of choice of contribution for the first three rounds of play; in two treatments all players are constrained to contribute €2 per round, while in the remaining two treatments half of the group has to contribute €4 per round and the other half is bound to €0 per round. Additionally, in two treatments (one with symmetric and one with unequal giving in the first three rounds), the subjects can pledge future contributions before beginning round four and also before the last three rounds. The central hypothesis advanced in the paper is that the real-world features introduced in the climate game, namely inequalities among actors and the ability to make non-binding pledges, have deep consequences on the cooperation level. Both claims

that the inequality disrupts cooperation and the pledges help coordination are supported by the data. 70% of the groups provided the public good in the symmetric treatment with pledges, relative to 50% in the corresponding treatment without pledges; 60% successful cooperation obtained in the asymmetric treatment with pledges, while only 20% obtained in the corresponding treatment without pledges.

Estratto per riassunto della tesi di dottorato

Studente: Alessandro Tavoni matricola: 955272

Dottorato: Economia

Ciclo: 22°

Titolo della tesi : Essays on Fairness Heuristics and Environmental Dilemmas

Abstract: The issues explored in this work concern individual behaviour and its departure from the rationality paradigm. While different in terms of underlying methodology, the chapters share the unifying theme of fairness as a guiding principle for human behaviour, as well as a focus on its relevance for environmental dilemmas.

Estratto: Le questioni affrontate nella tesi riguardano i comportamenti individuali e i relativi scostamenti dal paradigma di razionalità. Nonostante l'utilizzo di metodologie diverse nei tre capitoli, essi hanno in comune il tema unificante di equità come principio guida del comportamento umano, così come una particolare attenzione alla sua rilevanza nei dilemmi ambientali.