Moral Rules and Social Preferences in Co-operation Problems

BY ERNESTO M. GAVASSA PÉREZ*

DATE: 23RD OF MARCH 2022

****Note:** This is a preliminary version of the paper submitted for the 8th annual workshop on behavioural game theory hosted by the University of East Anglia. A suggested citation for the work presented herein is: Gavassa Pérez, E. (2022). Moral Rules and Social Preferences in Cooperation Problems. In E. M. Gavassa Pérez, *From Morality to Rules to Choices: Introducing and Testing a New Theory on how Morals Influence Cooperation* (pp. 77-125). Nottingham: PhD Thesis.**

Abstract: In this paper I introduce the MRC framework, which presents two theories about how moral rules drive cooperative behaviour: blame avoidance, or an imperative to avoid doing what one considers as blameworthy, and praise seeking, or an imperative to do what one considers as praiseworthy. Using this new framework, I test the extent to which these two moral rules and a set of preference-based theories (selfishness, inequality aversion, reciprocity, spite, social efficiency and maximin) can explain people's attitudes to cooperation in two co-operation problems: social dilemmas, where the individual and social optima are misaligned, and common interest games, where the individual and social optima are aligned. My results suggest that (i) blame avoidance, inequality aversion and maximin preferences are the best candidate explanations of people's attitudes to cooperation in both problems; (ii) praise seeking, reciprocity, social efficiency, and selfishness are also explanations of attitudes to cooperation in common interest games; and (iii) spite is the least promising explanation of attitudes to cooperation in either co-operation problem.

JEL codes: B12, B40, C91, D01, D91, H41

Keywords: Common Interest Games; Cooperation; Inequality Aversion; Maximin; Morality; Public Goods; Reciprocity; Social Dilemmas; Social Efficiency; Social Preferences; Spite; Spite Dilemma.

*Gavassa Pérez: University of Nottingham, School of Economics, Sir Clive Granger Building, University Park, Nottingham, NG7 2RD, United Kingdom (e-mail: ernesto.gavassaperez@nottingham.ac.uk). This paper is part of the PhD Thesis of Ernesto María Gavassa Pérez. Support from the School of Economics in the form of a PhD Scholarship, and of CeDEx for funding the experiment, is gratefully acknowledged. The author also thanks his mother for the emotional support, and Robin Cubitt, Simon Gächter, Konstantinos Ioannidis, Alex Possajennikov, Martin Sefton, Vernon Smith, and Bertil Tungodden for very constructive comments on the experimental design and methodology that we hereby present.

1. Introduction

The objective of my paper is to study whether, and if so how, moral judgments and social preferences influence contribution attitudes in two public goods problems: a social dilemma game, where individual and social interests are opposed, and a common interest game, where individual and social interests are aligned. Throughout this paper I define contribution attitudes as the schedule of preferred contributions, for different average contribution levels of other members of the group.

To achieve this, I elicit each subject's moral judgments of all strategy combinations of both public goods problems, and I present a new framework, the MRC framework, that uses such moral judgments to make predictions of contribution attitudes in both public goods problems. We introduce two moral rules within the MRC framework (each of them providing us with a different prediction for a subject's contribution attitudes): *blame avoidance*, or an imperative to avoid doing blameworthy actions, and *praise seeking*, or an imperative to do the most praiseworthy actions. Additionally, I use several experimental games to elicit, at the individual level, the parameters of a set of social preference models (inequality aversion, maximin, reciprocity, social efficiency, and spite); and use the elicited parameters to calculate, for each subject and social preference, each subject's optimal contribution attitudes in both public goods problems. By eliciting for each experimental subject the contribution attitudes in both co-operation problems, and comparing them to the predictions of the social preference and moral rules models, I can observe the predictive success of all the considered theories at the individual level and establish which of their underlying motivational factors are determinants of contribution attitudes in social dilemma games and common interest games.

Public goods are ubiquitous in human social life. We vote to maintain democracy, and we appreciate traffic rules and primary education, among other goods, daily. Yet, we cannot exclude other members of a community from using those goods if they do not contribute to them. The neoclassical economics framework, assuming strictly selfish individuals, predicts the under provision of public goods (see Samuelson, 1954). However, there exist some '*privileged groups*' where at least some – if not all – of its members find it profitable to fully contribute at the individual level to provide the public good (see Olson, 1965, pp. 49-50). Experiments in the 1970's onwards reported that, in one-shot interactions, subjects significantly deviated from the theoretical predictions by contributing around half of their endowment in social dilemmas (see Ledyard, 1995, and Zelmer, 2003 for reviews), and they also deviated by contributing less than optimally in common interest games (see Saijo and

Nakamura, 1995)[1]. To rationalize these behaviours, economists challenged the assumption of the selfish utility and allowed different social motives to be included within a subject's utility (see Sobel, 2005 and Cooper and Kagel, 2017)[2]. More recent research shows that people's attitudes to contribution are such that many people tend to contribute more the higher the average contributions of other co-players, whereas a non-negligible share of subjects are free riders in social dilemmas (see Chaudhuri, 2011 for a review, and Fischbacher, Gächter and Fehr, 2001, and Fischbacher and Gächter, 2010)[3].

Despite the wide range of social preferences that can explain contribution attitudes (see, for instance, Fehr, Fischbacher and Gächter, 2002 and Fehr and Schmidt, 2006, pp. 669-673), explicit tests of the success of social preferences in predicting contribution attitudes are scarce, let alone i) tests that compare several theories at the same time, and ii) tests that analyze a theory's predictive success at the individual level (but see, for instance, Beranek et al, 2017 for a within-subjects test of inequality aversion's predictive power of contribution attitudes in a social dilemma game). Although in general social preferences showcase a high predictive success at the aggregate level, one of their flaws is their lower consistency at the individual level (see Blanco et al, 2011). In this paper, we contribute to the understanding of the underlying motivations behind contribution attitudes in public goods games by testing, at the individual level, several social preferences and two new moral rules, and investigate whether the latter are better predictors of contribution attitudes at the individual level.

Additionally, by examining jointly the contribution attitudes in social dilemmas and common interest games allows a theoretical separation between the predictions of the considered theories[4]. More specifically, we set our experimental design so that most social preferences (inequality aversion, reciprocity, social efficiency, and spite) could not predict a joint pattern of contribution attitudes that we conjectured, ex ante, to be prevalent among subjects (conditional co-operation in the social dilemma and unconditional co-operation in the common

[1] See Bohm (1972); Dawes, McTavish and Shaklee (1977), Marwell and Ames (1979) and Isaac Walker, and Thomas (1984) for early evidence on contributions to social dilemmas; and see also Palfrey and Prisbey (1997), Brunton et al (2001), Brandts et al (2004), and Reuben and Riedl (2009) for evidence on common interest games.

[2] For empirical evidence, see, as well, Andreoni (1988, 1990 and 1995); Croson (1996); Ferraro and Vossler (2010); Palfrey and Prisbey (1996 and 1997); Anderson, Goeree and Holt (1998). For theoretical models built to accommodate this evidence, see see Fehr and Schmidt (1999) and Bolton and Ockelfels (2000) for inequality aversion motives, Sugden (1984), Rabin (1993), Dufwenberg and Kirchsteiger (2004), and Cox et al (2007) for reciprocity motives, Falk and Fischbacher (2006) for a mixture of inequality aversion and reciprocity motives, Charness and Rabin (2002) for a mixture of social efficiency and maximin motives, Batigalli and Dufwenberg (2007) for guilt aversion motives, McKelvey and Palfrey (1995) for confusion motives, Cappelen et al (2007) for egalitarian, libertarian and liberal egalitarian concerns, Andreoni (1990) for impure altruistic concerns, and Levine (1998) for spiteful concerns.

[3] For literature on contribution attitudes to public goods, one can additionally refer to Weimann (1994), Bardsley (2000), Keser and Van Winden (2000), Frey and Meier (2004), Croson et al (2005), Herrmann and Thöni (2009), Neugebauer et al (2009), Smith (2011), Cartwright and Lovett (2014), Hartig et al (2015), Gächter et al (2017), Andreozzi et al (2020), and Eichenseer and Moser (2000) among others.

[4] This is highlighted in Palfrey and Prisbey (1997, see especially the discussion in pp. 830-831). But switching the focus to contribution attitudes makes the theoretical separation more interesting, as it allows me to differentiate between different social preference models (see section 4 and the supplementary material).

interest game). Hence, only maximin and the two moral rule theories within the MRC framework were ex-ante compatible with the conjectured pattern of joint contribution attitudes. Besides the theoretical usefulness of studying common interest games, they represent real life co-operation problems where parties have their interests aligned. Different public goods have different levels of productivity, and/or different intrinsic utility to agents. Hence, public goods with a high enough level of productivity or intrinsic utility for the agents in a community will resemble the common interest situation (see Olson, 1965 and Reuben and Riedl, 2009 for a discussion).

Another novelty of the paper is the development of a novel framework to model the influence of moral judgments in subjects' choices, inspired by the works of Sen (1977), Smith and Wilson (2019), and some moral philosophers[5]. Morality has been studied since ancient times and has been a way to prescribe different ways to act that were deemed good. Throughout history, moral philosophers have emphasized it as a motivational factor in people (e.g., see, for instance, David Hume's, 1960 quote '*morals excite passions, and produce or prevent actions'*). My framework departs from social preference models in two main ways. First, the MRC framework conjectures that it is *people's conscious normative evaluations* of positive concepts that explains people's actions. In short, it is not because 'this action yields unequal outcomes' why a person acts to avoid inequality. Rather, I propose that it is because this action yields unequal outcomes, and '*yielding unequal outcomes is immoral*', is the reason why a person actively refrains from choosing that action. Second, the MRC framework departs from a self-centered conception of decision making as it considers the moral judgments made from an impartial spectator stance to be the ones influencing a person's moral code of conduct. Whereas models of inequality aversion or reciprocity consider only inequality or reciprocity with respect to oneself, the MRC framework considers the moral judgment of a given strategy from a position where a person is detached from his/her stakes in the situation.

[5] Works that have influenced my view on the topic and prosociality and driven me to study morality are those of Aristotle (2004), Thomas Hobbes (1996 and 2008), the Earl of Shaftesbury (2000), Francis Hutcheson (2004 and 2004), David Hume (1960 and 1983), Adam Smith (1982), Kant (1998), Rousseau (1979), and John Stuart Mill (1998). In economics, there is another branch of the literature that tries to incorporate morality as a special case of a social preference function – see, most notably, Alger and Weibull (2013), and more recently Masclet and Dickinson (2019). One can additionally refer to Sen (1977), Tungodden (2004), or Vanberg (2015) for good discussions on the relation between economics and morality. The works, in economics, of Harsanyi (1955), Laffont (1975), Etzioni (1987), Bordignon (1990), Binmore (1998), Brekke et al (2003), Bilodeau and Gravel (2004), Bénabou and Tirole (2006), Croson (2007), Roemer (2010), Alm and Torgler (2011), Bénabou and Tirole (2011), Nielsen and Mcgregor (2013), Hodgson (2014), Blasch and Ohndorf (2015), Hauge (2015), Daube and Ulph (2016), Capraro and Rand (2018) and Friedland and Cole (2019) and the works, in psychology, of Blasi (1984), Kohlberg and Candee (1984), De Waal (1996), Nucci (1996), Fischer and Ravizza (2000), Aquino and Reed (2002), Fiske (2002), Hardy and Carlo (2005), Krebs and Denton (2005), Haidt (2008), Janoff-Bulman et al (2009), Rai and Fiske (2011), Ellemers and Van den Bos (2012), Fiske (2012), Gray et al (2012), Ellemers et al (2013), Curry (2016), Schein and Gray (2018), and Anderson et al (2020) among others serve to highlight the importance of morality in the literature of decision theory as a regulator of behaviour.

The statistical analysis of the experimental games indicates that attitudes to contribution in the social dilemma and the common interest game differ markedly. Whilst most people are either conditional co-operators or free riders in the social dilemma, a substantial number of subjects are unconditional co-operators in the common interest game, and the share of conditional co-operation in the common interest game is substantially lower. Interestingly, the unconditional co-operators in the common interest game are not the free riders in social dilemmas. Rather, most unconditional co-operators in the common interest game tend to be conditional co-operators in the social dilemma (as we conjectured). Additionally, I find that both moral judgments and social preferences determine people's contribution attitudes in both games. More specifically, blame avoidance, maximin, and inequality aversion motives are the major determinants of contribution attitudes in social dilemmas and common interest games. Reciprocity, social efficiency, praise seeking, and material selfishness are only determinants of contribution attitudes in common interest games, and spite is not a determinant of contribution attitudes in either co-operation problem.

The paper proceeds as follows. Section 2 presents the experimental design. Section 3 presents the novel theoretical framework and its theoretical predictions. Section 4 discusses the theoretical predictions of the social preference models I consider. Section 5 presents the results of my experiment and section 6 concludes.

2. Experimental design

Each subject completed eight experimental tasks. Three of them – an *ultimatum game* (henceforth, UG), and a set of *modified dictator games* (henceforth, MDG) and *reciprocity games* (henceforth, RG) – were designed to elicit the parameters of a set of social preferences. Two experimental tasks involved two different versions of a two-person, one-shot, simultaneous move public goods game. I refer to these versions as a *social dilemma game* (henceforth, SDG) and a *common interest game* (henceforth, CIG), and to the tasks related to these versions as *P-experiments*. They elicited each subject's *contribution attitudes (*as defined above – a subject's desired schedule of contributions for each contribution of the other group member*)*. Additionally, subjects had to complete what I refer to as two *M-experiments*, one related to the SDG and another related to the CIG. The M-experiments elicited each subject's moral judgments of each strategy combination of the SD and the CIG. Finally, subjects also completed a sociodemographic questionnaire.

For the remainder of the paper, I refer to all tasks related to the SDG (the relevant P- and M-experiments) as the *social dilemma tasks* and to all tasks related to the CIG (the relevant P- and M-experiments) as the *common interest game tasks*. I also refer to tasks involving UG, MDG and RG as *parameter-elicitation tasks*.

The order in which subjects performed the experimental tasks was as follows. Everyone answered the sociodemographic questionnaire at the end and the parameter-elicitation tasks after all the social dilemma and common interest game tasks had been completed. The sequence in which all subjects answered the parameter elicitation tasks was kept the same for all: they completed the UG first, followed by the RG and, finally, the MDG. In contrast, I manipulated two aspects of the order of tasks: (i) whether the social dilemma tasks preceded or followed the common interest game tasks; and (ii) whether the M-experiments preceded or followed the P-experiments. This led to four different sequences in which tasks could be presented, which I outline in the Appendix, Table A1.

This manipulation led to a mixed design, where each subject had to complete all the tasks (*within-subjects* component) and subjects were randomly assigned to a treatment arm with a particular sequence (*between-subjects* component). The rationale for this design choice is threefold. First, moral suasion in public goods has been documented previously (see Dal Bó and Dal Bó, 2014). I wanted to control for any spillover effects between the M-experiments and the P-experiments to clearly identify any relation between moral judgments and contribution attitudes beyond that captured by order effects in the presentation of the tasks. Second, I wanted to control for spillover effects between social dilemma tasks and common interest game tasks. Since they are very similar games, I want to be sure I can control for any anchoring effect that may arise by having been exposed to a similar game before when analyzing contribution attitudes. Third, by eliciting the P-experiments, M-experiments, and the parameters for each subject I was able to get each subject's observed contribution attitudes of the SDG and the CIG and the predictions that each of the considered models make for those contribution attitudes. The within-subjects element of the design allowed us, thus, to have all the necessary information to test the theories at the individual level.

To ensure that subjects understood the incentives of the SDG and the CIG, they had to answer some control questions after reading the instructions but before completing the M- and P-experiments. Only after they answered all control questions correctly they could proceed to complete those tasks. Subjects were allowed to participate in the experiment once only, and they received no feedback on their earnings and co-player's decisions until all tasks had been completed. This procedure is similar to that of Blanco et al (2011) and minimizes the chance of learning about the co-player's choices between tasks.

Only the two P-experiments and the parameter-elicitation games were incentivized. The incentivization scheme was as follows. Subjects played different games, each game had different roles and two games (RG and MDG) had different versions with different payoff allocations. I first gathered all the data, and, at the end of the experiment, I randomly assigned subjects to games, and all subjects assigned to a given game were randomly matched into pairs. Once subjects were matched into pairs, I randomly assigned each pair member to one of the two possible roles for the game they had been allocated to. Lastly, for games with several versions (RG and MDG) one of the versions was randomly chosen to be relevant for each pair. Only the relevant actions arising from the randomization procedure implemented determined our subjects' final payoffs. Subjects were briefed about the procedure and knew how payoff were calculated. They also knew that all games, roles, and versions had the same probability of being chosen.

In the next subsections I provide a description of all tasks subjects had to complete. Given that one of the aims of the paper is to study the motivations behind contribution attitudes in social dilemmas and common interest games, I start by giving a detailed account of the public goods game I used in the experiment prior to briefly presenting each experimental task.

2.1. The public goods game

The two cooperation problems I study – SDG and CIG – are based on the same decision situation: a linear, one-shot, simultaneous move, two-person public goods game. In the public goods game versions I implemented, each of the group members is endowed with 30 tokens and must decide how many to contribute to a group project (the public good). The material payoff function of a generic subject i is:

$$(1) \qquad 30 - c_i + m * (c_i + c_{-i})$$

Where c_i (c_{-i}) refers to the token contributions of i (i's co-player) to the public good. A subject's feasible contribution levels are constrained to 0, 10, 20 or 30 tokens. For each token a subject does not contribute to the public good, that subject gets 1 token, and all the other group members get nothing. For each token a subject contributes to the public good, every member gets $m \in \{\underline{m}, \overline{m}\}$ tokens – that is, the benefits of the public good are non-excludable.

For the social dilemma I set $\underline{m}$ to 0.6, and for the common interest game I set $\overline{m}$ to 1.2[6]. Although the functional form of the payoff function is the same for both games, the qualitative incentive structure of the games is different because of the difference in the value of m. In the SDG, a subject gets more by not contributing a token to the public good (as $1 > 0.6$) whereas the total social payoff is maximized by contributing that token (as $1.2 > 1$). In contrast, in the CIG both the individual and total social payoff are maximized by contributing the token to the public good ($1.2 > 1$, and $2.4 > 1$ respectively).

2.2. Experimental tasks

2.2.1. The M-experiments

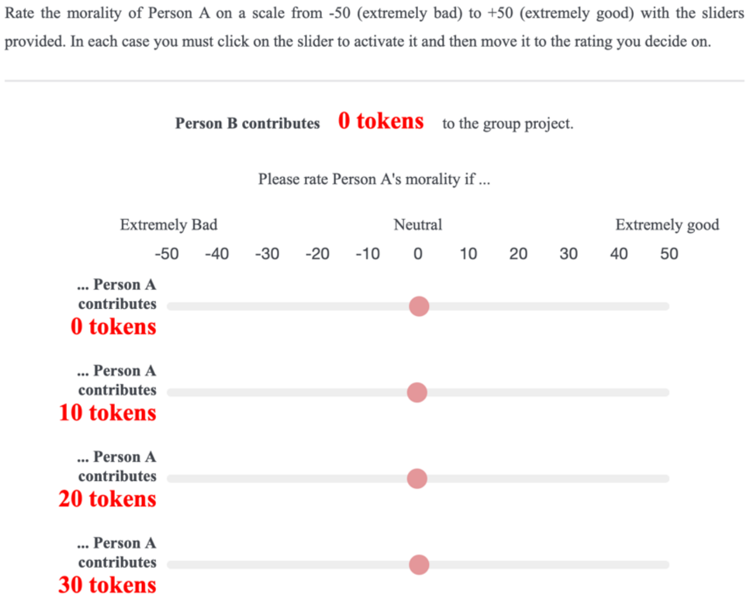
I use the survey method introduced by Cubitt et al (2011), and used in previous papers of the thesis, and adapt it to systematically elicit people's personal normative views of each strategy combination of the SD and the CIG.

Each M-experiment starts by presenting a given game to our subjects as an interaction between Person A and Person B. Then, I present each subject with several scenarios. Each scenario presents the contributions made by Person A and Person B to the public good and asks subjects to rate the morality of Person A on a scale ranging from -50 (extremely bad) to +50 (extremely good). A moral judgment of 0 is labelled as neutral. I run two M-experiments, one regarding the SDG and another one regarding the CIG. Each M-experiment consists of 16 scenarios, as I present to subjects one scenario for each strategy combination of Person A and Person B and the M-experiments are based on the SDG and CIG described earlier, where two players interact, each having only 4 feasible contribution levels (0, 10, 20, and 30). Figure 1a provides a screenshot of how a set of scenarios of the SDG were presented to subjects, with Person B's contribution held constant but Person A's contribution varied across the scenarios in a given set. Recall that it is always Person A who is being judged.

Three characteristics of the M-experiment are worthy of discussion. First, I told subjects that they are neither Person A nor Person B and, rather, they are giving their moral views as an outside observer (an *impartial spectator*). This design choice aims to capture impartiality in moral judgments typical of the moral theories, among others, of Adam Smith (see Konow, 2009, 2012 for discussion of the topic). A third party or a spectator has been used in the economics literature previously (see, for instance, Fehr and Fischbacher, 2004, for the use of

[6] More generally, for a SDG, then $\frac{1}{n} < \underline{m} < 1$ and for a CIG, then $\overline{m} > 1$

third parties and, more recently, Konow, 2009, Smith and Wilson, 2014, Cappelen et al, 2019 and Almas et al, 2020). It is because the theories I develop are based on the moral judgments that one forms as an impartial spectator guiding one's own behaviour that I implemented this design choice. Figure 1b summarizes how I introduced this feature to subjects in the M-experiment for the SDG.



Rate the morality of Person A on a scale from -50 (extremely bad) to +50 (extremely good) with the sliders provided. In each case you must click on the slider to activate it and then move it to the rating you decide on.

Person B contributes **0 tokens** to the group project.

Please rate Person A's morality if ...

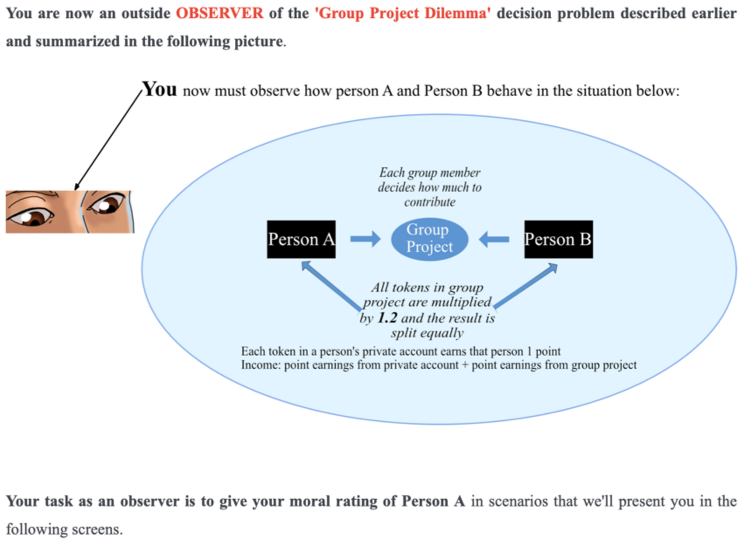
... Person A contributes 0 tokens

... Person A contributes 10 tokens

... Person A contributes 20 tokens

... Person A contributes 30 tokens

FIG 1A. SCREENSHOT OF SOME SCENARIOS OF THE SOCIAL DILEMMA'S M-EXPERIMENT



You are now an outside OBSERVER of the 'Group Project Dilemma' decision problem described earlier and summarized in the following picture.

You now must observe how person A and Person B behave in the situation below:

Each group member decides how much to contribute
Person A
Group Project
Person B
*All tokens in group project are multiplied by **1.2** and the result is split equally*
Each token in a person's private account earns that person 1 point
Income: point earnings from private account + point earnings from group project

Your task as an observer is to give your moral rating of Person A in scenarios that we'll present you in the following screens.

FIG 1B. IMPARTIAL SPECTATOR FEATURE OF THE M-EXPERIMENTS: IMPLEMENTATION

Second, subjects were explicitly told to give their own moral views rather than society's normative opinions about the scenarios. I use this approach as the theories I present in this paper are based on an individual's moral code rather than the social moral conventions. This follows the tradition of an important part of moral philosophy (see Russell, 2009, ch.42, p.334-344 for a discussion)[7].

Third, the M-experiments are not incentivized. I made this decision so that I did not confound subjects' true moral views with some hypothetical moral views that, if reported, would have maximized their payoff in the M-experiment given the incentive structure I would have chosen for it (see Cubitt at al, 2011 for discussion of this topic)[8]. This departs from what is currently done in the literature of social norms, where incentivized coordination games are used to elicit subjects' beliefs about the norms in their group (see Krupka and Weber, 2013 for one such approach). As good as this procedure sounds in the right context, it would not be appropriate for my design as I focus on subject's individual views rather than on their perceptions of the average social or moral conventions.

2.2.2. The P-experiments

I implement two tasks for both the SDG and the CIG: an *unconditional contribution* and a *contribution table task*. In the unconditional contribution task, a subject has to choose their contribution level without knowing what the other group member will choose. In the contribution table task, each subject must state their desired contribution per each feasible contribution of the other player. As each subject has four potential contribution levels (0, 10, 20, or 30), the contribution table task elicits four contributions per subject, one for each contribution level of the other player. It is this schedule of contributions from the contribution table task that I refer to as the subject's contribution attitudes, and which constitutes the dependent variable in our statistical analyses. Implementing the contribution table task in the SDG and CIG allows me to elicit such attitudes for both cooperation problems. The joint incentive-compatible elicitation of both tasks per each game constitutes the core methodology developed in Fischbacher, Gächter and Fehr (2001)[9], to which I refer to as the P-experiment.

To fix some notation, I define a free rider as a subject whose contributions are of the type $c_i^* = 0 \forall c_{-i}$; a perfect conditional cooperator as a subject whose contributions are of the type

[7] I do not wish to extend unnecessarily on this point, given the space constraints. But, as a very clear defence of this view see a claim of Russell's work cited in the main text: "*There are some who would say that a man need only obey the accepted moral code of his community. But I do not think any student of anthropology could be content with this answer. Such practices as cannibalism, human sacrifice, and head hunting have died out as a result of moral protests against conventional moral opinion.*"

[8] Additionally, there exists preliminary evidence suggesting that self-reported data contains important information aligning with subjects' attitudes in prosocial environments (see, for instance, Cappelen et al, 2011).

[9] To make both tasks incentive compatible, Fischbacher, Gächter and Fehr (2001) impose, to each group member, a probability p for the unconditional contribution task to be payoff relevant and a probability $1-p$ for the contribution table task to be payoff relevant. The probability p is known ex ante, but the realization of who will have the unconditional contribution and who will have the contribution table task as relevant is only realized after each subject has played both games.

$c_i^* = c_{-i} \forall c_{-i}$; and an unconditional cooperator as a subject whose contributions are of the type $c_i^* = 30 \forall c_{-i}$.

2.2.3. Parameter-elicitation games

Subjects played three different games to elicit the parameters of a set of social preference theories. One such game was the two-person, *ultimatum game*. In the generic ultimatum game (Güth et al, 1982), two players – a proposer and a responder – interact. In the first stage, the proposer's decision is the number of monetary units out of a total pie P to offer to the responder. In the second stage, the responder's decision is whether to accept the offer. Letting o denote the offer, the respondent's acceptance of the offer implies the proposer gets $P - o$ and the responder gets o units as payoff. If, however, the responder rejects the offer, both players get nothing. In essence, the respondent gets to decide between two allocations – $(P - o, o)$ and $(0,0)$ – where the first (last) entry in each of the allocations defines the proposer's (respondent's) material payoff. I impose the following restrictions to the parameters of the game: (i) $o \in \mathbb{N}^*$; (ii) $o \in \left[0, \frac{P}{2}\right]$, and iii) $P = 14$. Each subject had to make their decision as a proposer and decide whether to accept the offer for each potential o that the proposer can send.

I also presented to subjects a set of *modified dictator games* based on the ones described in Blanco et al (2011). In these games, the dictator must choose between keeping the full pie (denoted P, as before) for himself or split another pie $(2x)$ into two equal shares. In essence, it is a decision between two allocations – $(P, 0)$ and (x, x) – where the first (last) entry in each of the allocations defines the dictator's (recipient's) payoff. Implementing several versions of this game in which I keep P fixed and vary x allows me to elicit each subject's willingness to pay to implement an equal split of income. I impose the following restrictions when setting all the implementations of the game: i) $x \in \mathbb{N}^*$; ii) x is an even number; iii) $P = 20$; and iv) $x \in [0,32]$. Restriction iv) is a significant one as it allows subjects to reveal negative willingness to pay for implementing an equal split of the total pie for any $x > P$[10].

The *reciprocity games* I implemented followed the ones presented in Bruhin et al (2019). Each reciprocity game is a two-stage, sequential game. In the first stage, the first mover decides whether to implement the allocation – $(5,95)$ – or pass on that allocation. In the second stage, the second mover only gets to choose if the first mover passes from implementing $(5,95)$, in which case he can select one of two alternative allocations – (x_4, x_2) and $(0,0)$, where I only

[10] The direct implication is that, unlike Blanco et al (2011), I am explicitly able to detect subjects with spiteful preferences (i.e., subjects that derive pleasure for being ahead of others, and would need to be paid extra to accept an equal split of resources).

vary the alternative allocation (x_4, x_2) between versions of the reciprocity game. Across all reciprocity games, I impose $x_2 < 95$ so that the first mover's decision to pass on implementing the allocation $(5,95)$ is unambiguously unkind for the second mover (as either of the alternative distributions gives him/her a lower payoff). Each subject had to state, per each version, whether to pass on $(5,95)$ when playing the role of the first mover and which of the alternative allocations to select as the second mover.

I follow Blanco et al (2011) in using a revealed-preference approach based on the games just described to calibrate the parameters of all the social preference models I consider. Using this approach for all the choices made, the revealed-preference approach reveals a range of values for the relevant parameter – provided that the subject's responses are compatible with any (i.e., if choices do not violate any axiom underlying preference relations). In the supplementary material I present propositions showing the inequalities, for all the parameters of the social preference theories I consider, that are revealed given subjects' behaviour in the parameter elicitation games, but I briefly outline the intuition underlying the method in the following paragraphs.

As in Blanco et al (2011), I use the UG and the MDG to elicit the inequality aversion parameters (see Blanco et al, 2011 for a discussion on how to retrieve the inequality aversion parameters for each subject). Allowing for $x > P$ in the MDG allows me to capture negative values for the advantageous inequality parameter, which I use for a model of spiteful preferences. Additionally, the MDG allow me to extract the parameters of a social efficiency and a maximin model, and the reciprocity games allow me to retrieve the parameter of a model of sequential reciprocity.

I now briefly sketch the intuition behind the revealed preference approach for the social efficiency, maximin, and reciprocity models, starting with the social efficiency model. Whenever $x < P < 2x$, a subject's self-interest is better off with allocation $(P, 0)$ but a group's total payoff is better off when the subject chooses allocation (x, x). Hence, within the range $x \in \left[\frac{P}{2}, 20\right]$ there exists a tension between a subject's self-interest and social efficiency. The more money a subject is willing to forego (i.e., the higher $P - x$) to choose the equal allocation reveals a higher concern for social efficiency.

Regarding maximin, whenever $x \in [0,20] \leq P$ the person playing against the dictator will be worse off regardless of the allocation chosen (as $0 < P$, and $x \leq P$). Hence, within that range there will be a tension between increasing the payoff of the worse off by choosing (x, x) or maximizing one's own payoff by choosing $(P, 0)$. The more payoff a person is willing to forego (i.e., the higher $P - x$) to increase the payoff of the person worse off, the higher the concerns for maximin a person reveals to have.

Lastly, in the reciprocity games having chosen to pass on $(5,95)$ is perceived as unkind by the second mover, as $x_2 < 95$. Also, choosing the allocation $(0,0)$ instead of the allocation (x_4, x_2) is an unkind move towards the first mover, as $0 < x_4$. The higher the sum of money that the second mover is willing to forego (i.e., the higher the maximum x_2 rejected), the higher a subject's revealed willingness to reciprocate perceived unkindness with unkindness.

2.2.4. Sociodemographic questionnaire

Once subjects had finished all the previous tasks, I presented them several questions about their background characteristics. More specifically, I asked them about their gender, age, political identification (ranging from very left to very right), religiosity (ranging from not religious at all to very religious), the community size (in number of inhabitants) where they lived most of their life, their field of study and presented them with the big five personality traits questionnaire.

2.3. Participants and procedures

Due to Covid restrictions, I ran the experiment online during May 2021 using Qualtrics. I recruited 318 students from the University of Nottingham using the ORSEE platform (Greiner, 2015). The number of participants was determined by a power calculation aiming to achieve 80% power given available estimates from the previous paper (see the pre-registration document for more details). The average earning per subject being £7.88.

The average age of subjects was 21.4 years, 56.7% of subjects were female, another 51.9% identified as left and a further 42.5% self-reported as being religious. Subject choices in the Social Dilemma tasks (M- and P-experiments) were in line with the qualitative findings of previous papers. Additionally, the two different order manipulations did not significantly interact with my subjects' background characteristics and/or choices in the experimental tasks – see the supplementary material for an in-depth analysis of background characteristics and comparison with previously available data.

3. The MRC framework: from Morality to Rules to Choices

3.1. Motivation

The MRC framework models individuals as having impartial moral judgments (i.e., personal normative evaluations) of all strategy combinations of the decision situation of interest. It assumes that subjects have a moral rule that receives those moral judgments as inputs and outputs a set of normative prescriptions for desired play at the relevant decision situation. In the case there is more than one suggested way to proceed, material selfishness acts as a tiebreaker to decide which, among all the morally suggested actions, to choose. My methodological framework owes intellectually to the contribution of Smith and Wilson (2019), which transformed Adam Smith's moral theory into an economically tractable framework, and to Francis Hutcheson's (2004) and David Hume's (1960 and 1983) works. The framework I present is novel as it mixes some concepts of the latter philosophers to the general theory of Smith and Wilson (2019) to be able, for the first time, to use a theory of personal moral judgments to make precise, testable predictions of behaviour at the individual level.

The MRC framework departs from the classical way to model social preferences, which revolve around self-centered individuals pursuing the maximization of their own broadened utility, normally containing their material payoff along with a specific social goal (e.g., inequality aversion, reciprocity, social efficiency, maximin, spite, and so on). My framework, instead, is based on subjects whose impartial judgments influence the way they ought to act. There are three main points of departure with the classical way in which social preferences are modelled, which I proceed to discuss below.

Self-centeredness has been proven an undesirable feature of some of those models (i.e., models of direct reciprocity), as evidenced, for instance, by people's tendency to punish as third parties (see. most notably, Fehr and Fischbacher, 2004): it is because subjects cannot consider a harmful action geared towards another person as unkind why reciprocity cannot predict to engage in costly punishment as a third party. By modelling the way in which morality drives behaviour as impartial, I allow people to base their behaviour on how a situation is perceived regardless of whether it involves them.

Additionally, my framework assumes that it is not the properties of the social interaction that directly feed one's choice deliberation. Rather, it is subjects' implicit judgments about those properties that are relevant for their decisions: I assume that it is not because some outcomes are unequal why subjects avoid inequality; but, rather, that only if those unequal outcomes are morally blameworthy subjects will avoid them. Modelling morality in this way I allow subjects to act differently in payoff-equivalent situations to the extent that those situations are evaluated differently from a moral perspective, thereby allowing framing effects even when beliefs are held constant.

Lastly, as far as the suggestion from the moral rule is a unique choice, my framework assumes that it is only a subject's morality that drives their behaviour, rather than being a mixture of a social goal and material selfishness. This feature of morality as the only input to the decision-making process is a unique feature of the MRC framework and can capture deontological attitudes that have been widely documented in the moral psychology literature in the form of taboo-tradeoffs (for work on protected values, see Baron and Spranca, 1997 and Baron, 2017. For work on taboo trade-offs, see Tetlock 2003; Schoemaker and Tetlock, 2012; and Tetlock et al, 2017. For work on moral conviction, see Skitka et al, 2005; and Skitka, 2010. For work on morality as constraining the possible actions to be taken, see, more recently, Cushman, 2015; and Phillips and Cushman, 2017).

3.2. An illustrative example: the social dilemma game

To explain the intuition of my new framework, my starting point is the social dilemma game I presented in the previous section. Game theory typically assumes that a game is defined by the players, the set of strategies of each player and the utility functions of each player, that map each strategy combination into a given utility. Table 1a below presents the normal form matrix of the social dilemma game under the assumption that both players' utility depend exclusively on the material payoffs of the game. The row player is person i and the column player is i's opponent, which I name '$-i$'. Both players have free riding as a strictly dominating strategy, so the benchmark of material selfishness predicts free riding regardless of the contribution of the other player.

Table 1b transforms the material payoffs to account for inequality aversion as modelled by Fehr and Schmidt (1999). And, more generally, any social preference model changes this game theoretical benchmark by modifying the utility function of the players, thereby transforming the normal form matrix of material selfishness into a 'psychological' normal form matrix representing subjects' final utilities of every strategy combination of the game. In the case of inequality aversion, note that neither player will contribute more than the other player, as doing so decreases one's own material payoff and can only increase one's disadvantageous inequality, as $\alpha_i \geq 0$. However, inequality aversion deviates from the classical material selfishness assumption in the SDG whenever $\beta_i > 0.4$, as, in that case, each player's best response is to contribute the same as the other player ($c_i^* = c_{-i} \forall c_{-i} \in C$). Hence, inequality aversion can predict free riding or perfect conditional co-operation in the social dilemma game; and, crucially, the prediction will depend on the strength of a subject's aversion towards advantageous inequality.

Table 1. Normal form matrix of the SDG under material selfishness (A) and inequality aversion (B)

Normal form matrix of the Social Dilemma Game …

a. … assuming material self interest

$i \backslash -i$	$c_{-i} = 0$	$c_{-i} = 10$	$c_{-i} = 20$	$c_{-i} = 30$
$c_i = 0$	30,30	36,26	42,22	48,18
$c_i = 10$	26,36	32,32	38,28	44,24
$c_i = 20$	22,42	28,38	34,34	40,30
$c_i = 30$	18,48	24,44	30,40	36,36

b. … assuming Fehr-Schmidt preferences

$i \backslash -i$	$c_{-i} = 0$	$c_{-i} = 10$	$c_{-i} = 20$	$c_{-i} = 30$
$c_i = 0$	30,30	$36 - \beta_i 10{,}26 - \alpha_j 10$	$42 - \beta_i 20{,}22 - \alpha_j 20$	$48 - \beta_i 30{,}18 - \alpha_j 30$
$c_i = 10$	$26 - \alpha_i, 10{,}36 - \beta_j 10$	32,32	$38 - \beta_i 10{,}28 - \alpha_j 10$	$44 - \beta_i 20{,}24 - \alpha_j 20$
$c_i = 20$	$22 - \alpha_i 20{,}42 - \beta_j 20$	$28 - \alpha_i 10{,}38 - \beta_j 10$	34,34	$40 - \beta_i 10{,}30 - \alpha_j 10$
$c_i = 30$	$18 - \alpha_i 30{,}48 - \beta_j 30$	$24 - \alpha_i 20{,}44 - \beta_j 20$	$30 - \alpha_i 10{,}40 - \beta_j 10$	36,36

In contrast, the MRC framework elicits the moral judgments of every strategy combination in the social dilemma game, from an impartial perspective. Recall that moral judgments are on a scale from -50 (extremely bad) to +50 (extremely good). I represent such moral judgments in Table 2, setting the moral judgments to be the average moral judgments of the SDG in my experiments, rounded to the nearest integer, so that they are representative for the example.

Table 2. $\boldsymbol{i}$'s Moral judgments of Person A in the SDG

***i*'s Moral judgments of a Person A in the Social Dilemma Game …**

$a \backslash b$	$c_b = 0$	$c_b = 10$	$c_b = 20$	$c_b = 30$
$c_a = 0$	-3	-15	-25	-34
$c_a = 10$	$+12$	$+7$	-8	-17
$c_a = 20$	$+24$	$+20$	$+12$	-2
$c_a = 30$	$+37$	$+32$	$+29$	$+20$

The first evident difference with classical models of social preferences is that the matrix in Table 2 does not regard subject i, which is the focus of our attention. Social preferences are self-centered as they assume that i's worry about inequality is born out of how inequality influences him\her. Rather, the MRC framework contemplates morality as arising from a disinterested stance. To do this, I assume subject i rates the morality of a generic player, Person A, when playing against another generic player, Person B, in the same decision situation that person i will play. That is, the moral judgments of Person A are done in an environment where the set of strategies of Person A and Person B, and the payoff consequences of all strategy combinations, are the same as in the game that i plays against $-i$. The crucial assumption is that moral judgments are impartial. Thus, I assume that Person i will judge him/herself in the same way as he/she judges Person A. So, I can derive Table 3 from Table 2, where the moral judgments are kept the same, but now the players are i and $-i$.

TABLE 3. *i*'S MORAL JUDGMENTS OF HIM/HERSELF IN THE SDG

***i*'s Moral judgments of *i* in the Social Dilemma Game …**

$i \setminus -i$	$c_{-i} = 0$	$c_{-i} = 10$	$c_{-i} = 20$	$c_{-i} = 30$
$c_i = 0$	-3	-15	-25	-34
$c_i = 10$	$+12$	$+7$	-8	-17
$c_i = 20$	$+24$	$+20$	$+12$	-2
$c_i = 30$	$+37$	$+32$	$+29$	$+20$

The MRC assumes that the way subjects come to act is by following a moral rule. Following Smith and Wilson (2019), I propose two such rules within the MRC framework: blame avoidance and praise seeking. Both moral rules use the relevant moral judgments as inputs to produce a given choice, or set of choices, that are morally suggested.

Blame avoidance states that a person ought to avoid doing blameworthy actions (i.e., actions with negative moral judgments). In this example, then, blame avoidance suggests that a subject ought to avoid doing $c_i = 0$ against $c_{-i} = 0$, $c_i = 0$ against $c_{-i} = 10$, $c_i \in \{0,10\}$ against $c_{-i} = 20$ and $c_i \in \{0,10,20\}$ against $c_{-i} = 30$, as all are strategy combinations for which, by impartiality, I assume i will judge him/her as being blameworthy (i.e., with negative moral judgments).

Praise seeking states that a person ought to choose the most praiseworthy actions (i.e., actions with the highest moral judgment). Hence, this rule suggests that a person ought to choose $c_i = 30$ against $c_{-i} \in \{0,10,20,30\}$, as $c_i = 30$ has the highest rating attached to it for every value of c_{-i}.

In practice, these rules constrain the set of possible strategies to choose against each strategy combination, and I can represent their output with a modified Table 1a matrix in Tables 4a and 4b.

Table 4a represents the normal form matrix of the SDG with all the cells representing strategy combinations not suggested by blame avoidance shaded in grey. Similarly, Table 4b represents the normal form matrix of the SDG with all the cells representing strategy combinations not suggested by praise seeking shaded in grey. Cells shaded in grey are cells that cannot be chosen by an individual if he/she decides to follow the relevant moral rule (blame avoidance for table 4a; praise seeking for table 4b).

Table 4. Normal form matrix of the SDG under blame avoidance (a) and praise seeking (b)

Modified normal form matrix of the Social Dilemma Game …

a. … assuming blame avoidance

$i \setminus -i$	$c_{-i} = 0$	$c_{-i} = 10$	$c_{-i} = 20$	$c_{-i} = 30$
$c_i = 0$				
$c_i = 10$	26,36	32,32		
$c_i = 20$	22,42	28,38	34,34	
$c_i = 30$	18,48	24,44	30,40	36,36

b. … assuming praise seeking

$i \setminus -i$	$c_{-i} = 0$	$c_{-i} = 10$	$c_{-i} = 20$	$c_{-i} = 30$
$c_i = 0$				
$c_i = 10$				
$c_i = 20$				
$c_i = 30$	18,48	24,44	30,40	36,36

Whenever a moral rule suggests a single strategy to be taken, as is the case with praise seeking in Table 4b, then no further work is needed, and the relevant moral rule would predict those strategies to be chosen. In the case of praise seeking, it would imply that person i ought to be an unconditional co-operator (i.e., $c_i = 30 \forall c_{-i}$). If, however, more than one strategy is plausible given the output of a moral rule, as is the case with blame avoidance, then I use material selfishness as a tiebreaker to make a point prediction about i's play in the game. In the case of Table 4a, person i ought to choose $c_i = 10$ against $c_{-i} \in \{0,10\}$; choose $c_i = 20$ against $c_{-i} = 20$; and choose $c_i = 30$ against $c_{-i} = 30$.

3.3. A formal presentation of the MRC framework: praise seeking and blame avoidance

3.3.1. Preliminaries

Let $I \coloneqq \{i, -i\}$ be the set of players and $G \coloneqq \{SDG, CIG\}$, with g as its typical element, be the set of games; where SDG is the social dilemma and CIG is the common interest game. Let

$M \coloneqq \{-50, \dots.,0, \dots., +50\}$ be the judgment space. Let $C \coloneqq \{0,10,20,30\}$ be the individual contributions space in the public goods games presented earlier. It is the set of strategies (feasible contributions) for each hypothetical agent (Person A and Person B), for person i and for '$-i$'. Let the Cartesian product $C \times C$, with typical ordered pair $\langle c_a, c_b \rangle$, be the set of all strategy combinations in the public goods games I study; where c_a and c_b denote, respectively, the contributions of Person A (the judged person) and Person B (the non-judged person) to the public good. As $C \times C$ is also the set of strategy combinations of i and $-i$, I shall also use, without any loss of generality, the notation $\langle c_i, c_{-i} \rangle$ to refer to a typical ordered pair of $C \times C$. Let $m: C \times C \times G \times I \rightarrow M$ be the moral judgments of an impartial spectator of the set of the strategy combinations of the relevant games. Let, m depend on the strategy combination, the game being played and the identity of the person standing on the role of an impartial spectator: $m(\langle c_a, c_b \rangle, g, i)$. The variable i captures a subject i's biases that he/she cannot get rid of when entering the impartial spectator stance. Also, let $m_i: C \times C \times G \rightarrow M$ denote a function from the set of strategy combinations of relevant games to the judgment space. m_i is the function of the moral judgments that subject i holds about him/herself in game g for a strategy combination $\langle c_i, c_{-i} \rangle$. It follows that $m(\langle c_a, c_b \rangle, g, i) \in M$ represents the moral judgment that subject i has, as an impartial spectator, of Person A given the strategy combination $\langle c_a, c_b \rangle$ in game g. Similarly, $m_i(\langle c_i, c_{-i} \rangle, g) \in M$ represents the moral judgment that subject i has of him/herself given the strategy combination $\langle c_i, c_{-i} \rangle$ in game g. Lastly, denote $R: G \times C \rightarrow C$ as a function whose domain is all the combinations of strategies of a given player and relevant games and whose range is the set of strategies, common to all relevant games. Then, a function R can be understood as the rule that selects a given strategy against each strategy of the other player in each game. The functions of the type R, thus, represent the predicted schedules of contributions against each potential contribution of the other player in each game.

3.3.2. Assumptions of the MRC framework and predictions of blame avoidance and praise seeking

The MRC framework is based on five main assumptions: (1) impartiality in judgments; (2) subjectivity in judgments; (3) moral rules as constraints in choices; (4) material selfishness as a tiebreaker; and (5) rule-following. Below I present the five assumptions together with the predictions that blame avoidance and praise seeking make about contribution attitudes in the SDG and CIG. I discuss how each assumption is applied to both praise seeking and blame avoidance when the assumption is specific to each theory.

Assumption 1. Impartiality in judgments.

Assumption 1 says that subjects form moral judgments from the stance of an impartial spectator. Put differently, subjects evaluate the moral judgment of a given scenario imagining how they would judge such scenario if they would not take part in it. Then, they ascribe to themselves the same moral rating as they ascribed to the relevant player from the impartial spectator stance. This assumption is most prominent in Adam Smith's Theory of Moral Sentiments, but it also appears in other theories of moral philosophy, such as Hume's *judicious spectator* in the Treatise of Human Nature (1960 – Selby-Bigge edition, Book III, Part I, Sect. II., pp. 472), or Rawls' *veil of ignorance* within the original position proposed in A Theory of Justice (1999, pp.118-123). Given my notation, this assumption can be written as:

(2) $$If\ \langle c_a, c_b\rangle = \langle c_i, c_{-i}\rangle, then\ m(\langle c_a, c_b\rangle, g, i) \equiv m_i(\langle c_i, c_{-i}\rangle, g)$$

I use this assumption in the experiments to infer each subject's moral judgments of him/herself in all strategy combinations of the SDG and CIG from the moral judgments that they ascribed to Person A in the M-experiments (see discussion in subsection 3.2, where I go from Table 2 to Table 3). It is this assumption that makes the MRC framework to depart from the self-centeredness of classical models of social preferences, as I move the focus from analysing a social situation with respect to oneself (as social preferences do) to analysing the moral aspect of a scenario without subjects making any reference to themselves.

Assumption 2. Subjectivity in judgments.

Assumption 2 says that, although subjects put themselves in an impartial position when making judgments, nothing ensures that they can abstract from all their own characteristics when making judgments. Given my notation, I can capture Assumption 2 as:

(3) $$\frac{\partial m(\langle c_a, c_b\rangle, g, i)}{\partial i} \gtreqless 0$$

As far as the bias that two subjects bring to the impartial spectator stance is different, then their moral judgment of the same scenario will be different. In my notation,

(4) $If\ m(\langle c_a, c_b\rangle, g, i) \neq m(\langle c_a, c_b\rangle, g, -i), then\ m_i(\langle c_i, c_{-i}\rangle, g) \neq m_{-i}(\langle c_i, c_{-i}\rangle, g)$

Thus, Assumption 2's contribution to the MRC framework is to state that $m(\langle c_a, c_b\rangle, g, i) = m(\langle c_a, c_b\rangle, g, -i)$ is not necessarily true. This feature of moral judgments is especially present in the works of Francis Hutcheson (2002) and David Hume (1960), who held a view that paralleled aesthetics with ethics. They conceived that people may have different perceptions of *good* and *wrong*, just as they had different perceptions of *beauty* and *deformity*[11]. It is this assumption that makes the MRC framework different from Smith and Wilson (2019)'s Humanomics framework, as I consider subject's moral judgments – and, hence, potentially their predicted choices – to differ.

Assumption 3. Moral Rules as constraints to choices.

This assumption says that moral rules constrain the set of strategies to a subset of strategies that a subject can make in a game. I initially include two moral rules within the MRC framework: praise seeking and blame avoidance.

The rule of praise seeking states that subjects ought to seek choosing strategy combinations that they perceive as most praiseworthy as impartial spectators. Given my previous notation, I can define the subset of strategies suggested by the rule of praise seeking for individual i against strategy c_{-i} in game g as:

(5) $$B_{i,c_{-i},g} \coloneqq \{c_i \in C \mid (\forall c_i' \in C)\big(m_i(\langle c_i, c_{-i}\rangle, g) \geq m_i(\langle c_i', c_{-i}\rangle, g)\big)\}$$

Where B stands for 'best' and $B_{i,c_{-i},g} \subseteq C$ is the subset of strategies that praise seeking suggests an agent i to take against c_{-i} in game g. They are those strategies with the highest moral judgment for the relevant c_{-i} and g.

The *rule of blame avoidance* states that subjects ought to avoid choosing strategy combinations that they perceive as blameworthy as impartial spectators. Given my previous

[11] Read, for instance, Hume's (1998, pp.134) sentence: "There are certain terms in language which import blame, and others praise; and all men who use the same tongue must agree in their application of them. … But when critics come to particulars, this seeming unanimity vanishes; and it is found, that they had affixed a very different meaning to their expressions. … Those who found morality on sentiment, more than on reason, are inclined to comprehend ethics under the former observation, and to maintain, that, in all questions which regard to conduct and manners, the difference among men is really greater than at first sight it appears"

notation, I can define the subset of strategies suggested by the rule of blame avoidance for individual i against strategy c_{-i} in game g as:

$$(6) \qquad U_{i,c_{-i},g} \coloneqq \{c_i \in C | m_i(\langle c_i, c_{-i}\rangle, g) \geq 0\},$$

where U stands for 'un-condemned' and $U_{i,c_{-i},g} \subseteq C$ is the subset of strategies that blame avoidance suggests an agent i to take against c_{-i} in game g. These are those strategies that have a non-negative moral judgment for the relevant c_{-i} and g.

Assumption 4. Material selfishness as a tiebreaker.

This assumption says that with respect to their material payoffs subjects are strictly monotonous, locally insatiable individuals. Hence, in the absence of moral considerations they prefer to choose strategies that yield them a higher material payoff. In other words:

$$(7) \qquad (\forall c_i' \in C), c_i \succ c_i' \; iff \; \pi_i(\langle c_i, c_{-i}\rangle, g) > \pi_i(\langle c_i', c_{-i}\rangle, g)$$

Where $\pi_i(\langle c_i, c_{-i}\rangle, g)$ refers to the material payoff that subject i gets given the strategy combination $\langle c_i, c_{-i}\rangle$ in game g.

Whenever the sets $B_{i,c_{-i},g}$ or $U_{i,c_{-i},g}$ contain a single element, that is, $|B_{i,c_{-i},g}| = 1$ or $|U_{i,c_{-i},g}| = 1$ respectively, then subject i's choices against c_{-i} in game g will be uniquely determined by praise seeking or blame avoidance, respectively. However, whenever more than one strategy lies within $B_{i,c_{-i},g}$ or $U_{i,c_{-i},g}$, then I apply material selfishness as a tiebreaker to decide the predicted strategy for subject i against c_{-i} in game g. More formally,

$$(8) \qquad B'_{i,c_{-i},g} := \left\{ c_i \in B_{i,c_{-i},g} | \left(\forall c_i' \in B_{i,c_{-i},g}\right), c_i \succ c_i' \right\}$$

$$(9) \qquad U'_{i,c_{-i},g} := \left\{ c_i \in U_{i,c_{-i},g} | \left(\forall c_i' \in U_{i,c_{-i},g}\right), c_i \succ c_i' \right\}$$

Where the set $B'_{i,c_{-i},g} \subseteq B_{i,c_{-i},g}$ (resp. $U'_{i,c_{-i},g} \subseteq U_{i,c_{-i},g}$) represents a set with a single element, the element being the strategy that yields the highest payoff within all the strategies allowed by praise seeking (resp. blame avoidance) against c_{-i} in game g.

Assumption 5. *Rule-following.*

This assumption says that subjects make their choices according to their moral rules and, when the tiebreaker is needed, refined by material self-interest. The rules for praise seeking and blame avoidance for subject i when playing against c_{-i} in game g can be defined as:

$$PS_{i,c_{-i},g} \coloneqq \begin{cases} B_{i,c_{-i},g} \; if \; |B_{i,c_{-i},g}| = 1 \\ B'_{i,c_{-i},g} \; if \; |B_{i,c_{-i},g}| > 1 \end{cases} \tag{10}$$

$$BA_{i,c_{-i},g} \coloneqq \begin{cases} B'_{i,c_{-i},g} \; if \; U_{i,c_{-i},g} = \emptyset \\ U_{i,c_{-i},g} \; if \; |U_{i,c_{-i},g}| = 1 \\ U'_{i,c_{-i},g} \; if \; |U_{i,c_{-i},g}| > 1 \end{cases} \tag{11}$$

Where $PS_{i,c_{-i},g}$ (resp. $BA_{i,c_{-i},g}$) is a set with a single element, that element representing subject i's predicted strategy against c_{-i} in game g if i follows the rule of praise seeking (resp. blame avoidance). Whenever $B_{i,c_{-i},g}$ and $U_{i,c_{-i},g}$ contain a single element, then the values of the functions $PS_{i,c_{-i},g}$ and $BA_{i,c_{-i},g}$ are uniquely based on the moral constraints imposed on choice by blame avoidance and praise seeking. Whenever $B_{i,c_{-i},g}$ and $U_{i,c_{-i},g}$ contain more than one element, then the values of the functions $PS_{i,c_{-i},g}$ and $BA_{i,c_{-i},g}$ are based on the most selfish actions out of the ones allowed by praise seeking and blame avoidance. Whenever all moral judgments are negative, then $U_{i,c_{-i},g}$ will be empty, and hence a subject's suggestion will be to do that action which minimizes blameworthiness when performed. In the case where all feasible strategies are blameworthy, that suggestion will be the same as the one of praise seeking, as the strategy with the highest moral judgment will be the least negative one.

I can, then, use sets of the type $PS_{i,c_{-i},g}$ and $BA_{i,c_{-i},g}$ to define praise seeking and blame avoidance's predicted vector of contributions for subject i in game g as:

$$\overrightarrow{PS}_{i,g} = \left(PS_{i,0,g}, PS_{i,10,g}, PS_{i,20,g}, PS_{i,30,g}\right) \tag{12}$$

$$\overrightarrow{BA}_{i,g} = \left(BA_{i,0,g}, BA_{i,10,g}, BA_{i,20,g}, BA_{i,30,g}\right) \tag{13}$$

It is these two vectors per each subject i and per each game g that form the predictions of praise seeking and blame avoidance regarding contribution attitudes in the SDG and CIG.

4. Social preferences and contribution attitudes

In the previous section I presented the MRC framework, which introduced two moral rule theories (blame avoidance and praise seeking) and their predictions of contribution attitudes. In this section I present the intuition behind the theoretical predictions of contribution attitudes that the material selfishness, inequality aversion and sequential reciprocity models make, relegating the proofs to the supplementary material. Additionally, I present the other social preference models I use, but relegate all the discussion on their theoretical predictions of contribution attitudes to the supplementary material.

4.1. Material selfishness: Homo Economicus preferences

I start my theoretical discussion with the classical benchmark of material selfishness.

Proposition 1. If subject i maximizes the utility function $U_i^{HE}(c_i, c_{-i}) = \pi_i(c_i, c_{-i})$, where $\pi_i(c_i, c_{-i})$ denotes the material payoff of person i for the strategy combination in which i contributes c_i and the other player c_{-i}, subject i's optimal contributions will be $c_i^* = 0\ \forall c_{-i} \in C$ (resp. $c_i^* = 30\ \forall c_{-i} \in C$) in the SD (resp. CIG).

Intuition. The marginal utility of contributing is negative in the SDG and positive in the MDG. Hence, $c_i^* = 0\ \forall c_{-i} \in C$ (resp. $c_i^* = 30\ \forall c_{-i} \in C$) is the unique solution to subject i's maximization problem in the SD (resp. CIG)

4.2. Inequality Aversion: Fehr-Schmidt preferences

The first social preference model I consider is inequality aversion by Fehr and Schmidt (1999). The model is the result of two assumptions. First, a subject maximizes his or her own utility. Second, the subject's utility is formed by a linear combination of concerns for their own payoff and for inequality concerns. More specifically, for a two-person game the utility function of the model is specified by the following functional form:

$$(14) \qquad U_i^{FS}(\pi_i, \pi_{-i}) \coloneqq \pi_i - \alpha_i * Max\{\pi_{-i} - \pi_i, 0\} - \beta_i * Max\{\pi_i - \pi_{-i}, 0\}$$

Where π_i and π_{-i} denote the payoffs of subject i and the other subject in the interaction, and the parameters α_i and β_i represent the strength of subject i's aversions to disadvantageous and advantageous inequality respectively. The Fehr-Schmidt model imposes the following restrictions to the parameters: (i) $\alpha_i \geq \beta_i$; (ii) $\alpha_i, \beta_i \geq 0$; (iii) $\beta_i < 1$. These restrictions imply, respectively, that (i) disadvantageous inequality looms larger than advantageous inequality; (ii) inequality can never increase a subject's utility; (iii) a subject is unwilling to burn money to reduce advantageous inequality.

Proposition 2. If subject i maximizes the utility function $U_i^{FS}\big(\pi_i(c_i, c_{-i}), \pi_{-i}(c_i, c_{-i})\big)$, where i contributes c_i and the other player contributes c_{-i}, then subject i's contribution attitudes will be

(i), in the Social Dilemma,

(a) be a free rider ($c_i^* = 0\ \forall c_{-i} \in C$) iff $\beta_i < 1 - MPCR$
(b) be a perfect conditional co-operator ($c_i^* = c_{-i}\ \forall c_{-i} \in C$) iff $\beta_i > 1 - MPCR$
(c) be indifferent between $c_i \in [0, c_{-i}]$ iff $\beta_i = 1 - MPCR$

(ii), in the Common Interest Game,

(a) be an unconditional co-operator ($c_i^* = 30\ \forall c_{-i} \in C$) iff $\alpha_i < MPCR - 1$
(b) be a perfect conditional co-operator ($c_i^* = c_{-i}\ \forall c_{-i} \in C$) iff $\alpha_i > MPCR - 1$
(c) Be indifferent between $c_i \in [c_{-i}, 30]$ iff $\alpha_i = MPCR - 1$

Intuition. In either game, contributing the same as the other player gives equal material payoffs to both players. In the SDG, contributing more than others lowers one's own material payoff and increases disadvantageous inequality. Hence, no inequality averse player will do this. In contrast, in the SDG contributing less than the other player increases one's own material payoff at the expense of increasing advantageous inequality. Hence, only a player with a high aversion to advantageous inequality will forego their personal interest and increase their contributions to match that of the other player. Despite the same functional form of the payoff function, as now $\overline{m} > 1$, contributing to the public good in the CIG is individually profitable and free riding is against one's material self-interest. Hence, in the CIG contributing less than others lowers one's material payoff and increases advantageous inequality. It follows then that no inequality averse player will do this. In contrast, in the CIG contributing more than others increases one's own material payoff at the expense of increasing disadvantageous inequality. Hence, only a player with a high aversion to disadvantageous inequality will forego their personal interest and decrease their contributions to match that of the other player.

4.6. Reciprocity: Dufwenberg-Kirchsteiger preferences

The next social preference model I consider is sequential reciprocity by Dufwenberg and Kirchsteiger (2004). The model has two assumptions. First, a subject maximizes his or her own utility. Second, the subject's utility is formed by a linear combination of concerns for their own payoff and for reciprocity concerns. More specifically, for a two-person game the utility function of the model is specified by the following functional form:

(15) $$U_i^{\mathrm{DK}}(\pi_i, \pi_{-i}) \coloneqq \pi_i\left(a_i(h), b_{i,-i}(h)\right) + Y_{i,-i} * \kappa_{i,-i}\left(a_i(h), b_{i,-i}(h)\right) * \lambda_{i,-i,i}\left(b_{i,-i}(h), c_{i,-i,i}(h)\right)$$

Where π_i denotes the strategy of subject i, $Y_{i,-i}$ denotes subject i's strength of reciprocal concerns towards the other player, and $\kappa_{i,-i}$ and $\lambda_{i,-i,i}$ represent subject i's kindness and perceived kindness towards the other player respectively. $a_i(h)$ denotes player i's action at node h, $b_{i,-i}(h)$ denotes player i's first-order belief, updated at node h, about the other subject's play in the game and $c_{i,-i,i}(h)$ denotes player i's expectations about what the other player believes he/she'll do, updated at node h. I refer to $c_{i,-i,i}(h)$ as player i's second-order belief in node h. Subject i's kindness and perceived kindness functions are defined as in Dufwenberg and Kirchsteiger (2004). Put shortly, they depend on the concept of *equitable payoff*, defined as the average between the maximum payoff a player can give to another within all the strategies available to him/her and the minimum payoff a player can give to another within the set of all efficient strategies. Efficient strategies are the set of strategies for which there is no other strategy giving a higher payoff to at least one player and no lower payoff to the other players for any history of play and subsequent strategies.

Proposition 3. If subject i maximizes the utility function $U_i^{DK}\left(\pi_i(c_i, c_{-i}), \pi_{-i}(c_i, c_{-i})\right)$, where i contributes c_i, the other player contributes c_{-i}, and the other player moves first and subject i second, then subject i will (i), in the Social Dilemma,

(a) do $c_i^* = 0$ against $c_{-i} \in \{0{,}10\}$ regardless of $Y_{i,-i}$

(b) do $c_i^* = 0$ against $c_{-i} \in \{20{,}30\}$ iff $Y_{i,-i} < \frac{1-MPCR}{MPCR^2 \times (c_{-i}-15)}$

(c) do $c_i^* = 30$ against $c_{-i} \in \{20{,}30\}$ iff $Y_{i,-i} > \frac{1-MPCR}{MPCR^2 \times (c_{-i}-15)}$

(ii), in the Common Interest Game,

(d) do $c_i^* = 30$ against $c_{-i} = 30$ regardless of $Y_{i,-i}$

(e) do $c_i^* = 0$ against $c_{-i} \in \{0{,}10{,}20\})$ iff $Y_{ij} > \frac{MPCR-1}{MPCR^2 \times (30-c_{-i})}$

(f) do $c_i^* = 30$ against $c_{-i} \in \{0{,}10{,}20\})$ iff $Y_{ij} < \frac{MPCR-1}{MPCR^2 \times (30-c_{-i})}$

Intuition. In the social dilemma all strategies are efficient. In contrast, only full contribution in the common interest game is an efficient strategy as less than full contribution would give a lower payoff to all players. Hence, it follows that contributing half of one's endowment (full contribution) is the equitable payoff in the social dilemma (common interest game). This implies that contributions below (above) half of one's endowment will be perceived as unkind (kind) in the social dilemma, and that no contributions can be perceived as kind in the common interest game. In the social dilemma, being reciprocal against perceived unkind players is always optimal, as free riding is also the material payoff maximizing strategy. However, being reciprocal against perceived kind players generates a tension between reciprocal motives (being as kind as possible and fully contribute) and selfish motives (free riding). Only subjects with high enough concerns for reciprocity will reciprocate kind actions by fully contributing in the social dilemma; and all subjects will free ride against perceived unkind players in the social dilemma. In the common interest game, being unkind towards perceived unkind players implies free riding, which is opposite to the material payoff maximizing strategy in common interest games (full contribution). It, hence, follows that only subjects with high concerns for reciprocity will depart from unconditional co-operation in the common interest game.

4.3. Other social preference models: spitefulness, social efficiency and maximin

The three remaining models of social preferences that I use in my paper capture preferences for spite, social efficiency, and maximin and are captured, respectively, by the three following utility functions:

(16) $$U_i^{\mathrm{S}}(\pi_i, \pi_{-i}) \coloneqq \pi_i - \beta_i \times Max\{\pi_i - \pi_{-i}, 0\}$$

(17) $$U_i^{\mathrm{SE}}(\pi_i, \pi_{-i}) \coloneqq (1 - p_i)\pi_i + p_i \times (\pi_i + \pi_{-i})$$

(18) $$U_i^{\mathrm{MM}}(\pi_i, \pi_{-i}) \coloneqq (1 - q_i)\pi_i + q_i \times Min\{\pi_i, \pi_{-i}\}$$

where the spiteful model assumes $\beta_i \leq 0$, and the social efficiency and maximin models assume, respectively, $p_i \in [0,1]$ and $q_i \in [0,1]$.

5. Results

I start discussing some descriptive statistics of the main variables of my study (average moral judgments of the SDG and CIG, distribution of contribution attitudes in the SDG and CIG, and average values for the parameters of the relevant social preferences). I then continue by presenting the two main analyses of how blame avoidance, praise seeking and the set of social preferences I consider influence contribution attitudes in the SDG and CIG.

5.1. Descriptive statistics

5.1.1. Average moral judgments of co-operation problems

Figure 2 plots the average moral judgments (with 95% confidence intervals) of all scenarios of both M-experiments. I display average moral judgments in 4 panels, each panel containing all average moral judgments corresponding to scenarios based on the same contribution level of Person A (the judged Person. For short, c_a. For reference, c_a is displayed in the shaded box above each panel). I then arrange (within each panel) the average moral judgments according to what I call *Moral Evaluation Functions*. Based on Cubitt et al (2011) and the two previous papers, I define a *Moral Evaluation Function of* c_a (henceforth, MEF of c_a) as the average moral judgment that subjects ascribe to Person A, given that Person A contributes c_a, expressed as a function of the contribution of the non-judged Person (Person B. For short, c_b).

I display MEF's for the data of social dilemmas and common interest games. The horizontal and vertical axes are common to all panels, the former representing feasible values of c_b and the latter representing the moral rating of each average moral judgment. Moral ratings range, as explained earlier, from -50 (extremely bad) to +50 (extremely good), a moral rating of 0 being defined as of no moral significance. As a benchmark, I plot – in each panel – a black, dotted horizontal line at a moral rating of 0. This benchmark represents the MEF that, if observed, would indicate all scenarios to have no moral significance.

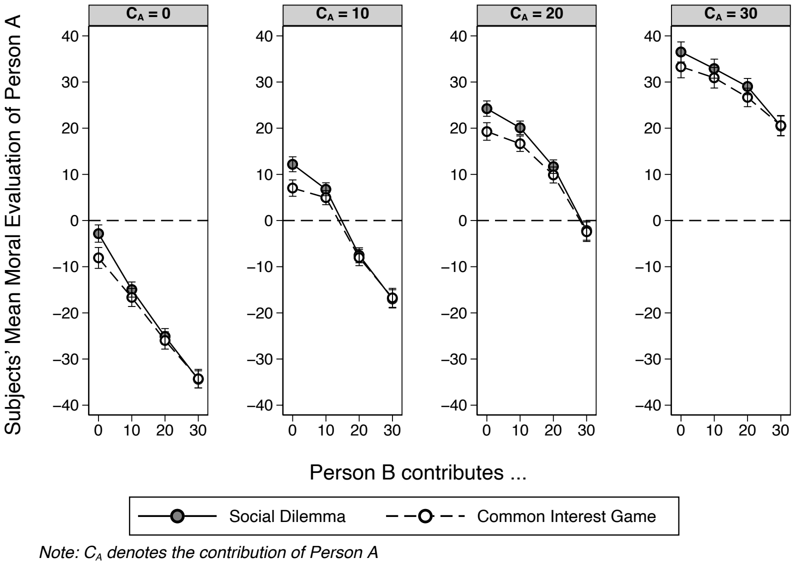


FIG 2. MORAL EVALUATION FUNCTIONS OF ALL CONTRIBUTIONS OF PERSON A.

Four features of Figure 2 are especially striking. First, average moral judgments different from 0 imply that subjects perceive the SDG and the CIG as situations of moral significance. Second, MEF's are increasing in c_a (the contribution of the judged person), suggesting an *increasing approbation of Person A* the more he/she contributes to the public good. Third, MEF's are decreasing in c_b (the contribution of the non-judged person), suggesting an *increasing condemnation of Person A* the higher the contribution of relevant others to the public good. And fourth – and perhaps more strikingly –, MEF's of social dilemmas and common interest games are remarkably similar. In practice, this means that subjects consider full contributions as morally equivalent in both games despite the non-sacrificial nature of full contribution in the CIG (i.e., contributions are individually profitable, so no material payoff sacrifice needs to be carried out to contribute to the public good in the CIG). Additionally, it means that free riding is perceived, on average, as morally equivalent despite its anti-social nature in the CIG (i.e., free riding can only generate advantageous inequality in the CIG).

I now discuss what the average moral judgments in Figure 2 reveal about the predicted contribution attitudes of praise seeking and blame avoidance in the SDG and CIG. Using the notation in Section 3, I can describe the predicted contribution attitudes of praise seeking as:

(19) $$\overrightarrow{PS}_{i,SDG} = \left(PS_{i,0,SDG}, PS_{i,10,SDG}, PS_{i,20,SDG}, PS_{i,30,SDG}\right) = (30,30,30,30)$$

(20) $$\overrightarrow{PS}_{i,CIG} = \left(PS_{i,0,CIG}, PS_{i,10,CIG}, PS_{i,20,CIG}, PS_{i,30,CIG}\right) = (30,30,30,30)$$

Fixing c_{-i} (the horizontal axis) at each of the four potential contribution levels in either game, reveals that full contribution is always perceived as the most praiseworthy action from an impartial spectator' point of view: praise seeking predicts unconditional co-operation. Regarding blame avoidance, I can describe its predicted contribution attitudes in the SDG and the CIG as:

$$(21) \qquad \overrightarrow{BA}_{i,SDG} = (BA_{i,0,SDG}, BA_{i,10,SDG}, BA_{i,20,SDG}, BA_{i,30,SDG}) = (10,10,20,30)$$

$$(22) \qquad \overrightarrow{BA}_{i,CIG} = (BA_{i,0,CIG}, BA_{i,10,CIG}, BA_{i,20,CIG}, BA_{i,30,CIG}) = (30,30,30,30)$$

Even though moral judgments are very similar in the SDG and CIG, blame avoidance makes different predictions for the SDG and CIG, which deserves some further comment. Since in the CIG full contribution is both the most selfish action and always has a non-negative moral rating, then unconditional contribution is blame avoidance's prediction in the CIG. In contrast, in the SDG the smaller the contribution the higher the material payoff. Hence, for each level of c_{-i} the smallest contribution level that has a non-negative moral rating will be blame avoidance's predicted contribution in the SDG. In Figure 2, these are $c_i = 10$ against $c_{-i} \in \{0,10\}$, $c_i = 20$ against $c_{-i} = 20$ and $c_i = 30$ against $c_{-i} = 30$; that is, conditional co-operation. This highlights an important feature of blame avoidance: it makes different predictions for different situations even when the observed moral judgments are equivalent across decision situations.

5.1.2. Contribution attitudes of co-operation problems

I now report in Table 5 the distribution of subjects' contribution attitudes in the SDG and CIG, which constitutes the dependent variable of my subsequent analyses. This analysis allows me to determine whether the distribution of contribution attitudes varies across games; and, incidentally, allows me to compare those observed contribution attitudes with the predicted contribution attitudes of praise seeking and blame avoidance given the moral judgments of Figure 2. I classify contribution attitudes in five types, according to the definitions provided in Thöni and Volk (2018): free riders, conditional co-operators, unconditional co-operators, hump-shaded and others.

TABLE 5. DISTRIBUTION OF CONTRIBUTION TYPES IN COOPERATION PROBLEMS.

	Social dilemma	Common interest game	χ^2	p value
Free riders	11.006%	0.943%	9.579	0.002
Unconditional co-operators	2.516%	33.648%	16.183	0.000
Conditional co-operators	80.189%	57.547%	14.235	0.000
Hump shaded	5.031%	3.459%	58.461	0.000
Other	1.258%	4.403%	20.012	0.000
Overall			119.218	0.000

The two most common contribution types in the social dilemma are conditional co-operators (approx. 80%) and free riders (approx. 11%). The predicted contribution attitude of Blame avoidance in the SDG is conditional co-operation, which makes blame avoidance, before any analysis, a good candidate to predict contribution attitudes in the SDG. In the common interest games, the two most common types are conditional co-operation (approx. 58%) and unconditional co-operation (approx. 34%). Nonparametric χ^2 tests show a statistically significant difference in the distribution of contribution types across games. More specifically, I find a significantly lower number of free riders and conditional co-operators and a significantly higher number of unconditional co-operators in the common interest game relative to the social dilemma. This switch from conditional co-operation to unconditional co-operation is predicted by the average contribution attitudes of blame avoidance, proving it as a good candidate to fit the data. Praise seeking, by predicting unconditional co-operation in both games, is *ex ante* better suited to be a determinant of contribution attitudes in the CIG.

I additionally report the joint distribution of types in Table 6. This analysis complements the previous one as it allows me to determine whether contribution attitudes vary within-subjects. I find the joint contribution of types as a very important measurement given that different social preferences favour different joint contribution types.

The data reveals that only three joint contribution types have a frequency of at least 5%. Conditional co-operation in both games is the most frequent joint contribution type (around 50% of subjects). Around 25% of subjects are conditional co-operators in the social dilemma and unconditional co-operators in the common interest game, and almost 6% of subjects are free riders in the social dilemma and unconditional co-operators in the common interest game. Around 44% of subjects have different contribution attitudes in the SDG and CIG, showing that for a substantial amount of the sample contribution attitudes are specific to the co-operation problem.

TABLE 6. JOINT DISTRIBUTION OF CONTRIBUTION TYPES IN COOPERATION PROBLEMS.

		Common interest game				
		Free riders	Unconditional co-operators	Conditional co-operators	Hump shaded	Other
Social dilemma game	Free riders	0.6%	6.0%	4.1%	0.0%	0.3%
	Unconditional co-operators	0.0%	2.5%	0.0%	0.0%	0.0%
	Conditional co-operators	0.3%	25.2%	50.3%	1.6%	2.8%
	Hump shaded	0.0%	0.0%	2.5%	1.9%	0.6%
	Other	0.0%	0.0%	0.6%	0.0%	0.6%

I find two patterns especially revealing. First, recall that unconditional contribution is the most selfish action in the CIG, as contributing to the public good gives a higher return than keeping tokens in one's private account $(1.2 > 1)$. Thus, if all unconditional co-operation were to come from selfish motives in the CIG, I would rather expect all the unconditional co-operators in the CIG to be free riders in the SDG. However, I observe that 75% of the unconditional co-operators in the common interest game are conditional co-operators in the social dilemma $(25.16/33.65 \approx 0.75)$, revealing that most unconditional co-operation in the CIG cannot born out of selfish concerns. Second, I designed the experiment so that, given the values of $\underline{m}$ and $\overline{m}$ that I chose, conditional co-operation in the SDG could not be compatible with unconditional co-operation in the CIG for inequality aversion and reciprocity. Also, social efficiency and spite are not compatible with conditional co-operation in the SDG. The high prevalence of conditional co-operators in social dilemmas and unconditional co-operators in the common interest game (approx. 25% of subjects) already suggests that a substantial amount of data can only be accounted by social preferences via maximin and by moral rules via blame avoidance.

5.1.3. Parameter estimates of social preference models

I end this subsection by presenting in Table 7 some descriptive statistics of elicited social preference parameters.

TABLE 7. ELICITED PARAMETERS

	Theoretical Range	Empirical range	25th percentile	Mean	75th percentile	St. dev.
Ineq. Aversion						
α_i	$[0, \infty)$	[0,3]	0.52	1.21	2.13	0.95
β_i	[0,1)	[0,1]	0.05	0.38	0.55	0.35
Spite						
β_i	$(-\infty, 0]$	$[-.61,0]$	0.00	-0.02	0.00	0.09
Reciprocity						
$Y_{i,-i}$	$[0, \infty)$	[0,3.9]	0.00	0.16	0.02	0.75
Social Efficiency						
p_i	[0,1]	[0,1]	0.06	0.47	1.00	0.43
Maximin						
q_i	[0,1]	[0,1]	0.05	0.38	0.55	0.35

Notes: The values of this table are computed without using the data of subjects with multiple switches in either of the three games. I maintain all remaining subjects regardless of whether they violate a condition of the theory (e.g., $\beta_i > \alpha_i$). For people with no switches, I impute values at the extreme of the theoretical range. For inequality aversion (resp. spite), I impute $\beta_i = 0$ whenever I observe $\beta_i < 0$ (resp. $\beta_i > 0$).

On average, the parameters of inequality aversion, social efficiency, and maximin are bigger than those of reciprocity and/or spite. In terms of behaviour, the average parameter values of inequality aversion and reciprocity imply free riding in SDG and a form of conditional co-operation in CIG. The average spite parameter is very close to 0 (-0.02), which implies the same predictions as material selfishness: free riding in social dilemmas and unconditional co-operation in common interest games. The average values of the social efficiency ($p_i = 0.47$) and the maximin ($q_i = 0.38$) parameters imply free riding in SDG and unconditional co-operation in CIG. Lastly, almost all parameters have a substantial standard deviation, and a mean outside the interquartile range in the spite and reciprocity parameters deserves some discussion. In the case of spite, most subjects in the modified dictator games elicited a positive β_i. I imputed a value of 0 for the spite parameter to all subjects who revealed a positive β_i, hence the skewed distribution. The distribution of the reciprocity parameter was also skewed as subjects showed extreme reciprocal attitudes in the reciprocity games: whereas around 62% of subjects revealed they preferred to burn no more than 15 units when the first mover passed on the distribution $(5,95)$, around 34% of subjects decided to burn more than 20 units to reciprocate the first mover's unkind action to pass on $(5,95)$. The high number of subjects with low revealed reciprocity dragged the mean downwards. In the supplementary material I report the histogram of all the elicited parameters[12].

[12] The observed joint distribution of the inequality aversion parameters replicates the qualitative features of previous studies (see Blanco et al, 2011, and Beranek et al, 2015): a non-negligible number of subjects (29%) violate the assumption $\alpha_i > \beta_i$ and I have a negative, significant spearman rank correlation between parameters ($\rho = -0.123; p$-value: 0.033).

5.2. Do social preferences and moral rules influence contribution attitudes of co-operation problems?

5.2.1. An econometric approach

I start my analysis by presenting random effects estimates of the data from the SDG and CIG separately. The equation I estimate uses the observed contribution attitudes as the dependent variable and the predicted contribution attitudes of most of the theories presented in the two previous sections as dependent variables[13]. Recall that contribution attitudes are elicited with the contribution table task on the P-experiments, which asks subjects to give a preferred contribution level against each potential contribution level of the other player. As the contribution space is restricted to $\{0,10,20,30\}$, this means that the contribution attitudes of a given subject in a game consist of four contributions, giving me a dependent variable with four observations per each subject for a given game. The predicted contribution attitudes of a given game of any theory consist of four observations as well: a predicted contribution per each observed contribution in the contribution table task. Whilst the predicted contribution attitudes of blame avoidance and praise seeking are calculated using the elicited moral judgments in the M-experiments (see Sections 3.2 and 3.3), the predicted contribution attitudes of the social preferences are calculated using the parameters elicited with the UG, the MDG and the RG. More specifically, I impute, for each subject, the theoretical best response (see the propositions in Section 4 and the Supplementary Material) given the parameter value elicited for him/her. I restrict the predictions to take the same potential values as the observed contributions. Additionally, in the estimated equation I also use the contribution of the other co-player (c_{-i}) to control for the potential effect of other relevant social preference theories in contribution attitudes, and two dummies to control for the order effects of moral judgments (whether moral judgments preceded or followed the P-experiment) and games (whether the SDG tasks preceded or followed the CIG tasks)[14]. Columns '*Estimates*' in Table 8 report the regression estimates.

[13] The regression analysis I report cannot include maximin preferences (spite) in the social dilemma as its predictions are the same as inequality aversion (the constant of regression). Additionally, I cannot include the predictions of social efficiency and maximin in the regression of common interest game. Again, this is due to the fact that their predictions are perfectly correlated with the constant of regression. The analysis of 5.2.2. includes all the 8 models in the comparison.

[14] My rationale is as follows. First, note that guilt aversion's prediction, in social dilemmas, of contribution attitudes for subjects with a high concern for avoiding guilt is contributing according to their second-order belief (see Dufwenberg et al, 2011). Assuming a high probability of playing against a conditional co-operator, it is reasonable to believe that the other co-player's contribution is increasing in that co-player's expectation about their contribution. Second, a central concept in social norms is empirical expectations (see Bicchieri, 2005 and 2017), which have been shown to be important drivers of behaviour even when they are in conflict with normative expectations (see Bicchieri and Xiao, 2009). As the contribution of others (c_{-i}) represents a subject's empirical expectations of his/her co-player behaviour I see a reasonable conjecture the statement that social norms' predictions will vary in proportion to c_{-i}.

Four patterns in the data reveal the role of each of the analysed theories in predicting contribution attitudes. First, only inequality aversion and blame avoidance are statistically significant in both games, which I take as a signal of them being more universal motives of contribution attitudes. Second, spite and social efficiency were statistically significant in the only regression in which they were included (CIG and SDG respectively). I take this as initial evidence of their role in explaining contribution attitudes. Third, reciprocity is statistically significant only in common interest games, suggesting that it is a specific motivation of contribution attitudes in the CIG. Four, only blame avoidance has a similar coefficient in both regressions, suggesting its effect is more stable than that of the other social preferences. More specifically, inequality aversion and reciprocity have a significantly greater coefficient in CIG, suggesting they play a greater role in explaining contribution attitudes in CIG.

TABLE 8. REGRESSION ESTIMATES AND DECOMPOSITION OF EXPLAINED VARIANCE.

Dependent variable: Contribution attitudes (elicited in the contribution table task of the P-experiments)

	Social dilemma game		*Common interest game*	
Independent variables	Estimates	Decomposition of R^2	Estimates	Decomposition of R^2
Constant	1.591		8.676***	
	(1.34)		(1.769)	
c_{-i}	0.585***	**52.58%**	0.213***	**20.77%**
	(0.031)		(0.05)	
Predictions				
Moral Rules				
Blame Avoidance	0.094***	**24.26%**	0.094***	**19.61%**
	(0.033)		(0.036)	
Praise Seeking	-0.011	0.40%	-0.021	0.33%
	(0.042)		(0.045)	
Social Preferences				
Inequality Aversion	0.11***	**17.46%**	0.225***	**34.34%**
	(0.034)		(0.044)	
Reciprocity	-0.006	0.71%	0.105***	**15.32%**
	(0.051)		(0.026)	
Social Efficiency	0.075**	4.07%		
	(0.03)			
Spite			0.06**	9.02%
			(0.026)	
Controls				
Social Dilemmas first	-0.596	0.24%	0.039	0.02%
	(0.743)		(0.931)	
Moral Judgments first	-0.873	0.28%	0.863	0.58%
	(0.741)		(0.929)	

Notes: * p<0.1 ** p<0.05 *** p<0.01. Percentages higher than 10% are printed in bold.

Additionally, I report the estimates of the decomposition of explained variance in columns '*Decomposition of* R^2' of Table 8. I decompose the explained overall variance in shares by applying the hierarchical partitioning method proposed in Chevan and Sutherland (1991) to the data. The share of all the independent variables adds up to one, each share representing the relative importance of each of the independent variables in explaining contribution attitudes.

It is remarkable to see that more than 50% of the explained variation in contribution attitudes of the SDG is captured by the c_{-i} control variable. As explained above, I used it as a proxy for the effect that other theories not included in the test had in contribution attitudes. More specifically, I conjectured guilt aversion and social norms to be the two main theories that could be represented within the control. The high relative importance in both games, together with statistical significance in both games, suggests that these alternative theories play an important role in contribution attitudes.

Going back to the theories I do actually test, blame avoidance appears as the clear winner in the SDG: its relative importance is higher than the aggregate relative importance of all the remaining theories (24.26% vs. 22.64%). Only inequality aversion gets close, capturing 17.46% of the explained variation of contribution attitudes in social dilemmas. Out of the remaining variables, only social efficiency has a non-negligible relative importance, although its role in explaining contribution attitudes is substantially lower than inequality aversion and blame avoidance.

Data from the CIG reveal a different picture, revealing inequality aversion as of greater relative importance than blame avoidance (34.34% vs 19.61%). Again, both theories share the first and second place of relative importance in the CIG. Reciprocity (15.32%) and spite (9.02%), this time, have a substantial degree of relative importance, strengthening my previous claim suggesting their game-specific role in explaining contribution attitudes of co-operation problems.

Overall, I observe three key messages revealed by the data. First, out of the theories tested only blame avoidance and inequality aversion are explanations of contribution attitudes in both co-operation problems. Second, reciprocity, social efficiency, and spite are game-specific explanations of contribution attitudes and play a minor role relative to that of blame avoidance and inequality aversion. Third, moral rules play a greater role than social preferences in explaining contribution attitudes of social dilemmas, and social preferences play a greater role than moral rules in explaining contribution attitudes of common interest games.

5.2.2. A revealed preference approach

I complement the econometric analysis of the previous subsection with an additional one because of one main concerns. Namely, I could not include all the theories in the econometric regressions since some theories made the same predictions, and some other theories were perfectly correlated with the constant of regression. Using a different approach, I can put to the test all the theories I consider in this paper against each other.

To solve the problem, I follow a revealed preference approach. Namely, I calculate some ratios that reveal the percentage of choices that i) reveal a given theory; and ii) reveal only that theory as compatible with the observed contribution attitudes in the SDG and CIG. I call those ratios the *degree of confirmation* and *degree of indubitable confirmation* of a theory by empirical evidence. I start by describing those ratios in detail before presenting the resulting data from the revealed preference approach.

5.2.2.1. Definitions of degree of confirmation and degree of indubitable confirmation

Let i denote an experimental subject, let g denote a game I investigate, let e_i^g denote the evidence provided by subject i in game g, and let t_i^g represent the theoretical predictions of theory t for experimental subject i in game g. Let I, G, E^g, and T^g be the sets containing, respectively, all relevant instances of i, g, e_i^g, and t_i^g. Then, I can define the degree of confirmation as the hit rate, or the relative frequency of successful predictions, that theory t makes in game g. Fixing n^g as the successful predictions of theory t in predicting the observed evidence of subject i (i.e., all instances of the type $t_i^g \equiv e_i^g$) and letting $N = |I|$ denote the cardinality of set I, or all the experimental subjects, I can write the *degree of confirmation* of theory t in game g given evidence E^g as:

$$C(t, E^g, g) = \frac{n^g}{N} \tag{23}$$

Fixing o^g to be the number of subjects for which **only** theory t successfully predicts the evidence (i.e., all instances of the type $t_i^g \equiv e_i^g$ where all rival theories r make predictions of the type $r_i^g \neq e_i^g$), I can write the *degree of indubitable confirmation* of theory t in game g given evidence E^g as:

$$I(t, E^g, g) = \frac{o^g}{N} \tag{24}$$

The rationale for using these two ratios to analyse the data is as follows[15]. The degree of confirmation, or the hit rate, of a given theory captures the share of the total data for which a

[15] One can also trace the use of hit rates as a way to capture the degree of confirmation of theories back to the philosophical tradition of logical empiricism. See, for instance, Reichenbach (1938, Ch. V, §39, pp.350-353) and Oppenheim (1945, pp.50). Also, see Popper (2002, part 2, paper 10, §79) for a critique of its use. Furthermore, one

given theory successfully predicts that data. Under a revealed preference approach, if option 1 is chosen when option 2 was available, then $U(1) > U(2)$ is inferred. As I have the theoretical predictions of each theory *ex ante*, I already know what option bears the highest utility for each theory. Hence, a choice compatible with the theoretical prediction reveals that a given theory's utility is revealed as compatible with the observed choice. By enumerating the share of observations compatible with each given theory I get a measurement of the total share of the data revealed to be compatible with a given theory. Also, by enumerating the share of observations that are only compatible with one of the theories (the degree of indubitable confirmation), I get a measurement of the total share of the observations that are revealed compatible with only one of the theories under test. I take this last measurement as the share of evidence that unambiguously favours that theory.

I make two further comments before I present the data. First, for a given theory to successfully predict the behaviour in a game (i.e., in my notation, all instances of the type $t_i^g \equiv e_i^g$) I impose that the full schedule of contribution attitudes must be correct. In other words, if a theory predicts correctly 3 out of 4 contributions in the contribution table task of a given game, that theory does not get a successful prediction for that individual. Second, I impose that a violation of an assumption of a given theory for an individual renders null any predictive power that the theory has. For example, if the calibrated parameters for individual i are $\beta_i = 0.7 > 0.5 = \alpha_i$, inequality aversion is not a successful prediction for subject i, as its preference parameters contradict one of the assumptions of the theory under test.

5.2.2.2. Estimates of degrees of confirmation and indubitable confirmation

I apply these two concepts to the data of SDG and CIG separately, and I additionally compute these ratios for the pooled data[16]. I present the estimates of $C(t, E^g, g)$ and $I(t, E^g, g)$ for all the theories I study in Table 9.

Looking at the data for SDG and CIG separately reveals blame avoidance as the clear winner of the analysis: not only has it a high degree of confirmation in both games but also, and more importantly, its degree of indubitable confirmation in each game is greater than the aggregate sum of that of all the alternative theories. Maximin can be declared to hold the second position in the contest as it displays a high degree of confirmation in both games and the second highest

can interpret the degree of indubitable confirmation we propose as a way to capture the share of what Bacon (2000, Book II, Aphorism XXXVI, pp.159-168) would call '*instantiae crucis*'.

[16] Given that the concepts are based on relative frequencies of successful predictions, and that each experimental subject provides evidence in both games, pooling the data means calculating successful instances over $2N$. Hence, I define the degrees of confirmation and indubitable confirmation of the pooled data, respectively, as $C(t,E) = \frac{n^{SD}+n^{CIG}}{2N}$ and $I(t,E) = \frac{o^{SD}+o^{CIG}}{2N}$, where the superscript $_{SD}$ $(_{CIG})$ refers to the version of the social dilemma (common interest game) I study in this paper.

degree of indubitable confirmation in SDG. Neither of the remaining theories display a degree of indubitable confirmation greater than 1%, but inequality aversion receives substantial degrees of confirmation (>15%) in both social dilemmas and common interest games. It is those three theories – blame avoidance, inequality aversion, and maximin – that I infer as the most probable explanations of contribution attitudes in SDG and CIG separately. Another subset of theories (reciprocity, social efficiency, material selfishness, and praise seeking) display a higher degree of confirmation in CIG than in SDG, and I infer them to be more likely explanations of contribution attitudes of CIG than of SDG. The evidence supports the inference of spite being decisively rejected as the explanation of contribution attitudes in SDG and CIG.

TABLE 9. OBSERVED DEGREES OF CONFIRMATION AND INDUBITABLE CONFIRMATION

	Social Dilemma		***Common Interest Game***		***Pooled***	
	$C(t, E^g, g)$	$I(t, E^g, g)$	$C(t, E^g, g)$	$I(t, E^g, g)$	$C(t, E)$	$I(t, E)$
Moral Rules						
Blame Avoidance	**26.42%**	**17.30%**	11.32%	**3.46%**	5.97%	4.09%
Praise Seeking	2.52%	0.94%	26.10%	0.00%	0.94%	0.31%
Homo Oeconomicus						
Selfishness	11.01%	0.00%	**33.02%**	0.00%	5.97%	0.31%
Social Preferences						
Inequality Aversion	17.92%	0.00%	18.87%	0.00%	7.23%	5.97%
Reciprocity	10.06%	0.00%	24.84%	0.00%	4.40%	0.00%
Social Efficiency	10.69%	0.63%	31.45%	0.00%	6.29%	0.94%
Maximin	**26.42%**	4.09%	31.45%	0.00%	**12.89%**	**6.92%**
Spite	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%

Notes: I print in bold the highest percentage in each column.

Another way to interpret the data is to assume that subjects are driven by a single motivation in all the situations they face. Looking at the degrees of confirmation and indubitable confirmation of the pooled data allows me to establish how each theory performs under this assumption. I consider this way of looking the data very important, given that a crucial motivation of running a within-subjects design was that theories made different predictions about the joint play in both games, and hence the within-subjects component allowed me to achieve a theoretical separation.

One word of caution when analysing the pooled data, though, is that the ratios I use are not *order-preserving* in a probabilistic sense. That is, the fact that a theory fares better than other in the SDG and CIG separately does not necessarily mean that such theory will also perform better than others in the pooled data. This is so as the data in the SDG and the CIG are potentially independent in nature. To see the point more intuitively, consider the following example: blame avoidance has a 50% degree of confirmation in the SDG and a 40% degree of confirmation in the CIG, and maximin has a 6% degree of confirmation in the SDG and a 5% degree of confirmation in the CIG. However, blame avoidance has 2% of instances where it successfully predicts the data of a subject in both the SDG and the CIG, whereas maximin has 4% of those instances. Hence, it follows that blame avoidance would have a pooled degree of confirmation of 2% whereas maximin would have a degree of confirmation of 4%, and maximin would have a higher pooled degree of confirmation even when blame avoidance has higher non-pooled degrees of confirmation.

Analysing the pooled data, maximin, inequality aversion, and blame avoidance are, again, the three best performing theories given that they display the highest pooled degrees of indubitable confirmation. What changes is the ranking of the three, being maximin the winner, and inequality aversion and blame avoidance holding, respectively, the second and third place. This is mainly because, at the individual level, maximin is compatible with more joint instances of conditional co-operation in the SDG and unconditional co-operation in the CIG than blame avoidance. Reciprocity, social efficiency, and material selfishness display similar pooled degrees of confirmation than the three winners but lower degrees of indubitable confirmation, showing a lower degree unambiguous evidence at the pooled level. Praise seeking and spite are the worst performing theories at the pooled level.

6. Concluding remarks and implications of the results

In this paper I have analysed the likelihood of a set of social preference and moral rule theories in explaining contribution attitudes of two co-operation problems: social dilemmas and common interest games. To achieve this, I have measured (i) contribution attitudes with P-experiments; (ii) the parameters of several social preference models with parameter-elicitation games; and (iii) the moral judgments of each strategy combination of social dilemmas and common interest games with M-experiments. The latter two measurements have been used to generate predictions of five social preference models (inequality aversion, reciprocity, social efficiency, maximin, and spite) and two novel moral rule models (blame avoidance and praise seeking). Using these theoretical predictions, I have tested the seven theories against each other, and against the benchmark of material selfishness, to determine the likelihood of each of them as explanations of contribution attitudes in co-operation problems.

I began my enquiry using econometric methods, which established the low likelihood of observing the data I observed if the null hypothesis (no theory explains contribution attitudes) were to be true. In addition, I was able to decompose the results into each theory's share of explained variation of contribution attitudes, finding that blame avoidance and inequality aversion held the higher shares of explained variation in both games. I took this as preliminary evidence supporting further investigation.

To provide a different insight into my results, I complemented the econometric analysis with a revealed preference approach, which allowed me to observe the degrees of confirmation and indubitable confirmation that each of the theories received from the data. The results agreed qualitatively with my previous findings and can be best summarized as follows. Within the inductive logic framework, I can group the theories into three clusters according to the confirmation they receive from the evidence. The first cluster, formed by maximin, inequality aversion, and blame avoidance, receives substantial confirmation as explanations of behaviour in both co-operation problems. The second cluster, including social efficiency, reciprocity, praise seeking, and material selfishness, only receive substantial confirmation in common interest games. The third cluster, formed by spite, contains the theories that receive no confirmation of an effect on contribution attitudes. In conclusion, contribution attitudes of co-operation problems are likely to be driven by several heterogeneous motivations.

One should be aware when interpretating the results, as there are a couple of potential objections that one can make to the claims I present above. First, by the way I elicit the parameters of social preferences, I imposed a consistency between the behaviour of both the SDG and the CIG on the one hand, and the behaviour in the parameter-elicitation games. In contrast, the moral rules have not required this consistency. While this is a plausible critique, one can still argue that the moral rules I present are required to match the data from the M-experiments with the contribution attitudes in the P-experiments, whereas none of the social preferences need to display such consistency. Hence, in my view, the different consistency requirements between the social preference theories and the moral rule theories just reflect the inherent differences between those theories.

Second, the falsification exercise relates to quantitative versions of the theories and not to qualitative ones. For instance, I have not allowed $\beta_i > \alpha_i$ in the Fehr and Schmidt (1999) model and, like Dufwenberg and Kirchsteiger (2004), I have not distinguished between positive and negative reciprocity. Hence, the failures of those models – or of any of the other social preferences I consider – can be related to any of the ancillary conditions of the test and not to a failure of the core concept of the theories (e.g., inequality aversion, reciprocity, and so on). Whilst this is true, it is an inherent feature of any experimental design to be subject to a Duhem-Quine thesis. In this specific case, I opted for a quantitative falsification as qualitative falsification of some concepts is virtually impossible. To see this, note that Rabin's (1993) reciprocity theory would predict free riding in both games, Sugden's (1984) reciprocity theory would predict perfect conditional co-operation in the SDG and CIG and Dufwenberg and Kirchsteiger's (2004) theory predicts either non-perfect conditional co-operation in both games or free riding in the SDG and non-perfect conditional co-operation in the CIG or free riding in the SDG and unconditional co-operation in the CIG. If one adds nonlinearity to the reciprocity element, as in Cox et al (2007), one would get a different sort of conditional co-operation. Thus, if one considers all theories related to a given concept one can end up in a situation where a given concept can predict every, or nearly every, possible behavioural pattern in a game. This would make the falsification exercise irrelevant and the theories pseudo-scientific, as they do not allow for behavioural patterns to contradict them. Hence, I have opted to choose specific versions of models that represented a given concept and that generated different predictions from other theories. In that vein, I chose Fehr and Schmidt (1999) as a way to capture perfect conditional co-operation in both games, Dufwenberg and Kirchsteiger (2004) to capture non-perfect conditional co-operation in both games, maximin to capture perfect conditional co-operation in the SDG and unconditional co-operation in the CIG, social efficiency to capture unconditional co-operation in both the SDG and the CIG, and spite to capture free riding in the SDG and conditional co-operation in the CIG in people's strengths for the social goal is strong enough. In this way, I was able to achieve theoretical separation between the behavioural content different concepts.

The results of this paper have two major implications that I proceed to discuss in detail now. One implication is that no unique motivation – at least from the ones considered in this study – can explain people's contribution attitudes. A second implication, more important in my view, is that the data does not support a single modelling strategy for representing subjects' social behaviour. The main modelling strategy in the social preferences literature relies on self-centered agents that derive pleasure from both material selfishness and a social goal. In contrast, the two moral rules within the MRC framework – praise seeking and blame avoidance – are models that represent an individual's motivation for the social as coming from a disinterested, impartial perspective. It is the individual's proactive judgment of the morality of the different scenarios that can arise in a decision situation that shape the content of the moral rules he/she is motivated to follow. This study demonstrates that both the classical, self-centered models and my new, impartial, moral judgment-based models are compatible with observed behaviour when the other models aren't, revealing two different paths to shaping contribution attitudes in social dilemmas and common interest games. Perhaps more interestingly, my study shows that blame avoidance, inequality aversion and maximin are the three theories with a higher degree of cross-game consistency, or within-subject predictive power. Whether the new framework is also able to inform a wider range of prosocial phenomena, like trust, gift-exchange, dictator giving, and ultimatum rejection of unequal offers among others, or cross-cultural or gender variation in behavioural traits, is an interesting task for future research.

7. REFERENCES

Alger, Ingela and Jörgen W. Weibull. 2013. "Homo Moralis—Preference Evolution under Incomplete Information and Assortative Matching." *Econometrica*, 81(6), 2269-302.

Alm, J. and B. Torgler. 2011. "Do Ethics Matter? Tax Compliance and Morality." *Journal of Business Ethics*, 101(4), 635-51.

Almås, Ingvild; Alexander W. Cappelen and Bertil Tungodden. 2020. "Cutthroat Capitalism Versus Cuddly Socialism: Are Americans More Meritocratic and Efficiency-Seeking Than Scandinavians?" *Journal of Political Economy*, 128(5), 1753-88.

Anderson, Rajen A.; Molly J. Crockett and David A. Pizarro. 2020. "A Theory of Moral Praise." *Trends in Cognitive Sciences*.

Anderson, Simon P.; Jacob K. Goeree and Charles A. Holt. 1998. "A Theoretical Analysis of Altruism and Decision Error in Public Goods Games." *Journal of Public Economics*, 70(2), 297-323.

Andreoni, James. 1995. "Cooperation in Public-Goods Experiments: Kindness or Confusion?" *The American Economic Review*, 85(4), 891-904.

____. 1990. "Impure Altruism and Donations to Public Goods: A Theory of Warm-Glow Giving." *The Economic Journal*, 100(401), 464-77.

____. 1988. "Why Free Ride?: Strategies and Learning in Public Goods Experiments." *Journal of Public Economics*, 37(3), 291-304.

Andreozzi, Luciano; Matteo Ploner and Ali Seyhun Saral. 2020. "The Stability of Conditional Cooperation: Beliefs Alone Cannot Explain the Decline of Cooperation in Social Dilemmas." *Scientific Reports*, 10(1), 13610.

Aquino, K. and A. Reed. 2002. "The Self-Importance of Moral Identity." *Journal of Personality and Social Psychology*, 83(6), 1423-40.

Aristotle. 2000. *Aristotle: Nicomachean Ethics*. Cambridge: Cambridge University Press.

Bacon, Francis. 2000. *The New Organon*. Cambridge: Cambridge University Press.

Bardsley, Nicholas. 2000. "Control without Deception: Individual Behaviour in Free-Riding Experiments Revisited." *Experimental Economics*, 3(3), 215-40.

Baron, Jonathan. 2017. "Protected Values and Other Types of Values." *Analyse & Kritik*, 39(1), 85-100.

Baron, Jonathan and Mark Spranca. 1997. "Protected Values." *Organizational Behavior and Human Decision Processes*, 70(1), 1-16.

Battigalli, Pierpaolo and Martin Dufwenberg. 2007. "Guilt in Games." *American Economic Review*, 97(2), 170-76.

Bénabou, Roland and Jean Tirole. 2011. "Identity, Morals, and Taboos: Beliefs as Assets *." *The Quarterly Journal of Economics*, 126(2), 805-55.

____. 2006. "Incentives and Prosocial Behavior." *American Economic Review*, 96(5), 1652-78.

Beranek, Benjamin; Robin Cubitt and Simon Gächter. 2017. "Does Inequality Aversion Explain Free Riding and Conditional Cooperation?," 1-63.

Bicchieri, Cristina. 2005. *The Grammar of Society: The Nature and Dynamics of Social Norms*. Cambridge: Cambridge University Press.

____. 2017. Norms in the Wild. How to Diagnose, Measure, and Change Social Norms. Corby: Oxford University Press.

Bicchieri, Cristina and Erte Xiao. 2009. "Do the Right Thing: But Only If Others Do So." *Journal of Behavioral Decision Making*, 22(2), 191-208.

Bilodeau, Marc and Nicolas Gravel. 2004. "Voluntary Provision of a Public Good and Individual Morality." *Journal of Public Economics*, 88(3), 645-66.

Binmore, Kenneth George. 1998. Game Theory and the Social Contract, Volume 2: Just Playing. United States: MIT press.

Blanco, Mariana; Dirk Engelmann and Hans Theo Normann. 2011. "A within-Subject Analysis of Other-Regarding Preferences." *Games and Economic Behavior*, 72(2), 321-38.

Blasch, Julia and Markus Ohndorf. 2015. "Altruism, Moral Norms and Social Approval: Joint Determinants of Individual Offset Behavior." *Ecological Economics*, 116, 251-60.

Blasi, A. 1984. "Moral Identity: Its Role in Moral Functioning," W. Kurtines and J. E. Gerwitz, *Morality, Moral Behaviour and Moral Development.* New York, United States of America: Wiley, 128-39.

Bohm, Peter. 1972. "Estimating Demand for Public Goods: An Experiment." *European Economic Review*, 3(2), 111-30.

Bolton, Gary E. and Axel Ockenfels. 2000. "Erc: A Theory of Equity, Reciprocity, and Competition." *American Economic Review*, 90(1), 166-93.

Bordignon, Massimo. 1990. "Was Kant Right?: Voluntary Provision of Public Goods under the Principle of Unconditional Commitment." *Economic Notes: Monte dei Paschi di Siena*, (3), 342-72.

Brandts, Jordi; Tatsuyoshi Saijo and Arthur Schram. 2004. "How Universal Is Behavior? A Four Country Comparison of Spite and Cooperation in Voluntary Contribution Mechanisms." *Public Choice*, 119(3), 381-424.

Brekke, Kjell Arne; Snorre Kverndokk and Karine Nyborg. 2003. "An Economic Model of Moral Motivation." *Journal of Public Economics*, 87(9), 1967-83.

Bruhin, Adrian; Ernst Fehr and Daniel Schunk. 2018. "The Many Faces of Human Sociality: Uncovering the Distribution and Stability of Social Preferences." *Journal of the European Economic Association*, 17(4), 1025-69.

Brunton, Douglas; Rabia Hasan and Stuart Mestelman. 2001. "The 'Spite' Dilemma: Spite or No Spite, Is There a Dilemma?" *Economics Letters*, 71(3), 405-12.

Cappelen, Alexander W.; Gauri Gauri and Bertil Tungodden. 2019. "Cooperation Creates Special Moral Obligations." *CESifo Working Paper*, No. 7052.

Cappelen, Alexander W.; Astri Drange Hole; Erik Ø Sørensen and Bertil Tungodden. 2011. "The Importance of Moral Reflection and Self-Reported Data in a Dictator Game with Production." *Social Choice and Welfare*, 36(1), 105-20.

____. 2007. "The Pluralism of Fairness Ideals: An Experimental Approach." *American Economic Review*, 97(3), 818-27.

Capraro, V. and D. G. Rand. 2018. "Do the Right Thing: Experimental Evidence That Preferences for Moral Behavior, Rather Than Equity or Efficiency Per Se, Drive Human Prosociality." *Judgment and Decision Making*, 13(1), 99-111.

Cartwright, Edward J. and Denise Lovett. 2014. "Conditional Cooperation and the Marginal Per Capita Return in Public Good Games." *Games*, 5(4), 234-56.

Charness, Gary and Matthew Rabin. 2002. "Understanding Social Preferences with Simple Tests*." *The Quarterly Journal of Economics*, 117(3), 817-69.

Chaudhuri, Ananish. 2011. "Sustaining Cooperation in Laboratory Public Goods Experiments: A Selective Survey of the Literature." *Experimental Economics*, 14(1), 47-83.

Chevan, Albert and Michael Sutherland. 1991. "Hierarchical Partitioning." *The American Statistician*, 45(2), 90-96.

Cooper, Davi J. and John H. Kagel. 2017. "Other-Regarding Preferences: A Selective Survey of Experimental Results," J. H. Kagel and A. E. Roth, *The Handbook of Experimental Economics, Volume 2.* Princeton, New Jersey: Princeton University Press, 217-89.

Cox, James C.; Daniel Friedman and Steven Gjerstad. 2007. "A Tractable Model of Reciprocity and Fairness." *Games and Economic Behavior*, 59(1), 17-45.

Croson, Rachel. 2007. "Theories of Commitment, Altruism and Reciprocity: Evidence from Linear Public Goods Games." *Economic Inquiry*, 45(2), 199-216.

Croson, Rachel; Enrique Fatas and Tibor Neugebauer. 2005. "Reciprocity, Matching and Conditional Cooperation in Two Public Goods Games." *Economics Letters*, 87(1), 95-101.

Croson, Rachel T. A. 1996. "Partners and Strangers Revisited." *Economics Letters*, 53(1), 25-32.

Cubitt, Robin P.; Michalis Drouvelis; Simon Gächter and Ruslan Kabalin. 2011. "Moral Judgments in Social Dilemmas: How Bad Is Free Riding?" *Journal of Public Economics*, 95(3), 253-64.

Curry, Oliver S. 2016. "Morality as Cooperation: A Problem-Centred Approach," T. K. Shackelford and R. D. Hansen, *The Evolution of Morality.* Switzerland: Springer, 27-51.

Cushman, Fiery. 2015. "From Moral Concern to Moral Constraint." *Current Opinion in Behavioral Sciences*, 3, 58-62.

Dal Bó, E. and P. Dal Bó. 2014. ""Do the Right Thing:" The Effects of Moral Suasion on Cooperation." *Journal of Public Economics*, 117, 28-38.

Daube, M. and D. Ulph. 2016. "Moral Behaviour, Altruism and Environmental Policy." *Environmental and Resource Economics*, 63(2), 505-22.

Dawes, Robyn M; Jeanne McTavish and Harriet Shaklee. 1977. "Behavior, Communication, and Assumptions About Other People's Behavior in a Commons Dilemma Situation." *Journal of Personality and Social Psychology*, 35(1), 1.

Dufwenberg, Martin; Simon Gächter and Heike Hennig-Schmidt. 2011. "The Framing of Games and the Psychology of Play." *Games and Economic Behavior*, 73(2), 459-78.

Dufwenberg, Martin and Georg Kirchsteiger. 2004. "A Theory of Sequential Reciprocity." *Games and Economic Behavior*, 47(2), 268-98.

Eichenseer, Michael and Johannes Moser. 2020. "Conditional Cooperation: Type Stability across Games." *Economics Letters*, 188, 108941.

Ellemers, Naomi; Stefano Pagliaro and Manuela Barreto. 2013. "Morality and Behavioural Regulation in Groups: A Social Identity Approach." *European Review of Social Psychology*, 24(1), 160-93.

Ellemers, Naomi and Kees van den Bos. 2012. "Morality in Groups: On the Social-Regulatory Functions of Right and Wrong." *Social and Personality Psychology Compass*, 6(12), 878-89.

Etzioni, Amitai. 1987. "Toward a Kantian Socio-Economics." *Review of Social Economy*, 45(1), 37-47.

Falk, Armin and Urs Fischbacher. 2006. "A Theory of Reciprocity." *Games and Economic Behavior*, 54(2), 293-315.

Fehr, Ernst and Urs Fischbacher. 2004. "Third-Party Punishment and Social Norms." *Evolution and Human Behavior*, 25(2), 63-87.

Fehr, Ernst; Urs Fischbacher and Simon Gächter. 2002. "Strong Reciprocity, Human Cooperation, and the Enforcement of Social Norms." *Human Nature*, 13(1), 1-25.

Fehr, Ernst and Klaus M. Schmidt. 2006. "The Economics of Fairness, Reciprocity and Altruism – Experimental Evidence and New Theories," S.-C. Kolm and J. M. Ythier, *Handbook of the Economics of Giving, Altruism and Reciprocity.* Elsevier, 615-91.

____. 1999. "A Theory of Fairness, Competition, and Cooperation*." *The Quarterly Journal of Economics*, 114(3), 817-68.

Ferraro, Paul J and Christian A Vossler. 2010. "The Source and Significance of Confusion in Public Goods Experiments." *The B.E. Journal of Economic Analysis & Policy*, 10(1).

Fischbacher, Urs and Simon Gächter. 2010. "Social Preferences, Beliefs, and the Dynamics of Free Riding in Public Goods Experiments." *American Economic Review*, 100(1), 541-56.

Fischbacher, Urs; Simon Gächter and Ernst Fehr. 2001. "Are People Conditionally Cooperative? Evidence from a Public Goods Experiment." *Economics Letters*, 71(3), 397-404.

Fischer, John Martin and Mark Ravizza. 2000. *Responsibility and Control: A Theory of Moral Responsibility*. United Kingdom: Cambridge university press.

Fiske, Alan Page. 2012. "Metarelational Models: Configurations of Social Relationships." *European Journal of Social Psychology*, 42(1), 2-18.

____. 2002. "Socio-Moral Emotions Motivate Action to Sustain Relationships." *Self and Identity*, 1(2), 169-75.

Frey, Bruno S. and Stephan Meier. 2004. "Social Comparisons and Pro-Social Behavior: Testing "Conditional Cooperation" in a Field Experiment." *American Economic Review*, 94(5), 1717-22.

Friedland, J. and B. M. Cole. 2019. "From Homo-Economicus to Homo-Virtus: A System-Theoretic Model for Raising Moral Self-Awareness." *Journal of Business Ethics*, 155(1), 191-205.

Gächter, Simon; Felix Kölle and Simone Quercia. 2017. "Reciprocity and the Tragedies of Maintaining and Providing the Commons." *Nature Human Behaviour*, 1(9), 650-56.

Gray, Kurt; Liane Young and Adam Waytz. 2012. "Mind Perception Is the Essence of Morality." *Psychological Inquiry*, 23(2), 101-24.

Greiner, Ben. 2015. "Subject Pool Recruitment Procedures: Organizing Experiments with Orsee." *Journal of the Economic Science Association*, 1(1), 114-25.

Güth, Werner; Rolf Schmittberger and Bernd Schwarze. 1982. "An Experimental Analysis of Ultimatum Bargaining." *Journal of Economic Behavior & Organization*, 3(4), 367-88.

Haidt, Jonathan. 2008. "Morality." *Perspectives on Psychological Science*, 3(1), 65-72.

Hardy, S. A. and G. Carlo. 2005. "Identity as a Source of Moral Motivation." *Human Development*, 48(4), 232-56.

Harsanyi, John C. 1955. "Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility." *Journal of Political Economy*, 63(4), 309-21.

Hartig, Björn; Bernd Irlenbusch and Felix Kölle. 2015. "Conditioning on What? Heterogeneous Contributions and Conditional Cooperation." *Journal of Behavioral and Experimental Economics*, 55, 48-64.

Hauge, Karen E. 2015. "Moral Opinions Are Conditional on the Behavior of Others." *Review of Social Economy*, 73(2), 154-75.

Helmer, Olaf and Paul Oppenheim. 1945. "A Syntactical Definition of Probability and of Degree of Confirmation." *Journal of Symbolic Logic*, 10(2), 25-60.

Herrmann, Benedikt and Christian Thöni. 2009. "Measuring Conditional Cooperation: A Replication Study in Russia." *Experimental Economics*, 12(1), 87-92.

Hobbes, Thomas. 2008. The Elements of Law Natural and Politic. Part I: Human Nature; Part Ii: De Corpore Politico. New York, United States of America: Oxford University Press.

Hobbes, Thomas and Richard Tuck. 1996. "Hobbes: Leviathan : Revised Student Edition."

Hodgson, Geoffrey M. 2014. "The Evolution of Morality and the End of Economic Man." *Journal of Evolutionary Economics*, 24(1), 83-106.

Hume, David. 1987. *An Enquiry Concerning the Principles of Morals*. Indianapolis: Hackett Pub. Co.

____. 2008. *Selected Essays*. Oxford: Oxford Univ. Press.

____. 1739. A Treatise of Human Nature: Being an Attempt to Introduce the Experimental Method of Reasoning into Moral Subjects. Oxford, United Kingdom: Clarendon Press.

Hutcheson, Francis. 2002. An Essay on the Nature and Conduct of the Passions and Affections, with Illustrations on the Moral Sense. Indianapolis, United Statis of America: Liberty Fund.

____. 2004. An Inquiry into the Original of Our Ideas of Beauty and Virtue in Two Treatises. Indianapolis, United States of America: Liberty Fund.

Isaac, R. Mark; James M. Walker and Susan H. Thomas. 1984. "Divergent Evidence on Free Riding: An Experimental Examination of Possible Explanations." *Public Choice*, 43(2), 113-49.

Janoff-Bulman, Ronnie; Sana Sheikh and Sebastian Hepp. 2009. "Proscriptive Versus Prescriptive Morality: Two Faces of Moral Regulation." *Journal of Personality and Social Psychology*, 96(3), 521-37.

Kant, Immanuel. 2012. *Groundwork of the Metaphysics of Morals*. Cambridge: Cambridge University Press.

Keser, Claudia and Frans Van Winden. 2000. "Conditional Cooperation and Voluntary Contributions to Public Goods." *The Scandinavian Journal of Economics*, 102(1), 23-39.

Kohlberg, Lawrence and Daniel Candee. 1984. "The Relationship of Moral Judgment to Moral Action," L. Kohlberg, *Essays in Moral Development: Vol. 2. The Psychology of Moral Development.* New York: Harper & Row, 498-581.

Konow, James. 2012. "Adam Smith and the Modern Science of Ethics." *Economics and Philosophy*, 28(3), 333-62.

____. 2009. "Is Fairness in the Eye of the Beholder? An Impartial Spectator Analysis of Justice." *Social Choice and Welfare*, 33(1), 101-27.

Krebs, D. L. and K. Denton. 2005. "Toward a More Pragmatic Approach to Morality: A Critical Evaluation of Kohlberg's Model." *Psychol Rev*, 112(3), 629-49.

Krupka, Erin L. and Roberto A. Weber. 2013. "Identifying Social Norms Using Coordination Games: Why Does Dictator Game Sharing Vary?" *Journal of the European Economic Association*, 11(3), 495-524.

Laffont, Jean-Jacques. 1975. "Macroeconomic Constraints, Economic Efficiency and Ethics: An Introduction to Kantian Economics." *Economica*, 42(168), 430-37.

Ledyard, John O. 1995. "Public Goods: A Survey of Experimental Research," J. H. Kagel and A. E. Roth, *The Handbook of Experimental Economics.* Princeton, New Jersey: Princeton University Press, 111-94.

Levine, David K. 1998. "Modeling Altruism and Spitefulness in Experiments." *Review of Economic Dynamics*, 1(3), 593-622.

Marwell, Gerald and Ruth E. Ames. 1979. "Experiments on the Provision of Public Goods. I. Resources, Interest, Group Size, and the Free-Rider Problem." *American Journal of Sociology*, 84(6), 1335-60.

Masclet, David and David L. Dickinson. 2019. "Incorporating Conditional Morality into Economic Decisions." *IZA Discussion Papers*, No. 12872.

McKelvey, Richard D. and Thomas R. Palfrey. 1995. "Quantal Response Equilibria for Normal Form Games." *Games and Economic Behavior*, 10(1), 6-38.

Mill, John Stuart. 1998. *Utilitarianism*. New York, United States of America: Oxford University Press.

Neugebauer, Tibor; Javier Perote; Ulrich Schmidt and Malte Loos. 2009. "Selfish-Biased Conditional Cooperation: On the Decline of Contributions in Repeated Public Goods Experiments." *Journal of Economic Psychology*, 30(1), 52-60.

Nielsen, L. and S. L. T. McGregor. 2013. "Consumer Morality and Moral Norms." *International Journal of Consumer Studies*, 37(5), 473-80.

Nucci, Larry P. 1996. "Morality and the Personal Sphere of Action.," E. S. Reed, E. Turiel and T. Brown, *Value and Knowledge.* New Jersey: Lawrence Erlbaum Associates, 41-60.

Olson, Mancur. 1965. *The Logic of Collective Action: Public Goods and the Theory of Groups*. Cambridge, Massachusetts: Harvard University Press.

Palfrey, Thomas R. and Jeffrey E. Prisbrey. 1996. "Altuism, Reputation and Noise in Linear Public Goods Experiments." *Journal of Public Economics*, 61(3), 409-27.

____. 1997. "Anomalous Behavior in Public Goods Experiments: How Much and Why?" *The American Economic Review*, 87(5), 829-46.

Phillips, Jonathan and Fiery Cushman. 2017. "Morality Constrains the Default Representation of What Is Possible." *Proceedings of the National Academy of Sciences*, 114(18), 4649-54.

Popper, Karl. 2002. *The Logic of Scientific Discovery*. London: Routledge Classics.

Rabin, Matthew. 1993. "Incorporating Fairness into Game Theory and Economics." *The American Economic Review*, 83(5), 1281-302.

Rai, Tage Shakti and Alan Page Fiske. 2011. "Moral Psychology Is Relationship Regulation: Moral Motives for Unity, Hierarchy, Equality, and Proportionality." *Psychological Review*, 118(1), 57-75.

Rawls, John. 1999. *A Theory of Justice. Revised Edition*. Cambridge, Massachusetts: The Belknap Press of Harvard University Press.

Reichenbach, Hans. 1938. *Experience and Prediction*. [Chicago]: University of Chicago Press.

Reuben, Ernesto and Arno Riedl. 2009. "Public Goods Provision and Sanctioning in Privileged Groups." *Journal of Conflict Resolution*, 53(1), 72-93.

Roemer, John E. 2010. "Kantian Equilibrium." *The Scandinavian Journal of Economics*, 112(1), 1-24.

Rousseau, Jean-Jacques. 1979. *Emile: Or on Education*. United States of America: Basic Books.

Russell, Bertrand. 2010. The Basic Writings of Bertrand Russell.

Saijo, Tatsuyoshi and Hideki Nakamura. 1995. "The “Spite” Dilemma in Voluntary Contribution Mechanism Experiments." *Journal of Conflict Resolution*, 39(3), 535-60.

Samuelson, Paul A. 1954. "The Pure Theory of Public Expenditure." *The Review of Economics and Statistics*, 36(4), 387-89.

Schein, Chelsea and Kurt Gray. 2018. "The Theory of Dyadic Morality: Reinventing Moral Judgment by Redefining Harm." *Personality and Social Psychology Review*, 22(1), 32-70.

Schoemaker, Paul J.H. and Philip E. Tetlock. 2012. "Taboo Scenarios: How to Think About the Unthinkable." *California Management Review*, 54(2), 5-24.

Sen, Amartya K. 1977. "Rational Fools: A Critique of the Behavioral Foundations of Economic Theory." *Philosophy & Public Affairs*, 6(4), 317-44.

Shaftesbury. 2000. *Characteristics of Men, Manners, Opinions, Times*. Cambridge University Press.

Skitka, Linda J. 2010. "The Psychology of Moral Conviction." *Social and Personality Psychology Compass*, 4(4), 267-81.

Skitka, Linda J.; Christopher W. Bauman and Edward G. Sargis. 2005. "Moral Conviction: Another Contributor to Attitude Strength or Something More?" *Journal of Personality and Social Psychology*, 88(6), 895-917.

Smith, Adam. 1982. *The Theory of Moral Sentiments*. Indianapolis: Liberty Classics.

Smith, Alexander. 2011. "Group Composition and Conditional Cooperation." *The Journal of Socio-Economics*, 40(5), 616-22.

Smith, Vernon L. and Bart J. Wilson. 2014. "Fair and Impartial Spectators in Experimental Economic Behavior." *Review of Behavioral Economics*, 1(1–2), 1-26.

____. 2019. Humanomics: Moral Sentiments and the Wealth of Nations for the Twenty-First Century. Cambridge: Cambridge University Press.

Sobel, Joel. 2005. "Interdependent Preferences and Reciprocity." *Journal of Economic Literature*, 43(2), 392-436.

Sugden, Robert. 1984. "Reciprocity: The Supply of Public Goods through Voluntary Contributions." *The Economic Journal*, 94(376), 772-87.

Tetlock, Philip E. 2003. "Thinking the Unthinkable: Sacred Values and Taboo Cognitions." *Trends in Cognitive Sciences*, 7(7), 320-24.

Tetlock, Philip E.; Barbara A. Mellers and J. Peter Scoblic. 2017. "Sacred Versus Pseudo-Sacred Values: How People Cope with Taboo Trade-Offs." *American Economic Review*, 107(5), 96-99.

Thöni, Christian and Stefan Volk. 2018. "Conditional Cooperation: Review and Refinement." *Economics Letters*, 171, 37-40.

Tungodden, Bertil. 2004. "Some Reflections on the Role of Moral Reasoning in Economics," NHH,

Vanberg, V. J. 2008. "On the Economics of Moral Preferences." *American Journal of Economics and Sociology*, 67(4), 605-28.

Waal, Frans B.M. de. 1997. *Good Natured. The Origins of Right and Wrong in Humans and Other Animals*. Cambridge, Massachusetts: Harvard University Press.

Weimann, Joachim. 1994. "Individual Behaviour in a Free Riding Experiment." *Journal of Public Economics*, 54(2), 185-200.

Zelmer, Jennifer. 2003. "Linear Public Goods Experiments: A Meta-Analysis." *Experimental Economics*, 6(3), 299-310.

8. Supplementary material: Theoretical derivations

8.1. Fixing some notation

The public goods game we consider is a 2-player, one-shot game. The relevant data from the P-experiment's strategy method (i.e., the conditional contribution task) is sequential in nature. To fix some notation before proceeding, we will henceforth refer to the two players in a group as player i and player j. We fix subject i's optimal contribution schedule in the conditional contribution task to be referred to as c_i^*; which will involve an optimal contribution against each potential contribution of the other player (that is, against each g_j). To make the notation more salient, and less prone to confusion with letter c, which we already use to denote the optimal contribution schedule, we opt to call a given contribution by player i as g_i, and a given contribution of player j by g_j. In mathematical terms, g_i and g_j are but generic contributions feasible for each player and lie within the sets $g_i \in A_i \coloneqq \{0,10,20,30\}$, and $g_j \in A_j \coloneqq \{0,10,20,30\}$. Hence, the cartesian product $A_i \times A_j$ refers to the set containing all strategy combinations of players i and j, and we denominate $\langle g_i, g_j \rangle$ (or, for notational compactness, g_i, g_j when within a parenthesis) to refer to a generic strategy combination of i and j that lie within the cartesian product defined earlier. The material payoff of player i (and analogously for player j) is represented by the following function:

$$\pi_i(g_i, g_j) = 30 - g_i + m \times (g_i + g_j)$$

Where $m \in \left(\frac{1}{n}, 1\right)$ for a social dilemma and $m \in (1, \infty)$ for a common interest game. At some points we will refer to $\underline{m}$ as an arbitrarily small value of the marginal per capita return and to $\overline{m}$ as an arbitrarily large value of the marginal per capita return to the public good. In all such instances, $\underline{m}$ will refer to a social dilemma game (that is, $\underline{m} \in \left(\frac{1}{n}, 1\right)$) and $\overline{m}$ will refer to a common interest game (that is, $\overline{m} \in (1, \infty)$).

8.2. The proofs

8.2.1. Predictions of theories regarding contribution preferences

8.2.1.1. An important lemma

For all the proofs that follow, and to shorten the derivations, we will use extensively a result. We summarise such a result in the following lemma:

Lemma 0. *In the aforementioned two-player, one-shot, public goods game, with the payoff functions* $\pi_i(g_i, g_j)$ *and* $\pi_j(g_i, g_j)$ *denoting, respectively, the payoffs of player* i *and player* j *from the strategy combination* $\langle g_i, g_j \rangle \in A_i \times A_j$*, it follows that:*

(a) $\pi_i(g_i, g_j) > \pi_j(g_i, g_j)\ iff\ g_i < g_j$.

(b) $\pi_i(g_i, g_j) - \pi_j(g_i, g_j) = g_j - g_i$ *and* $\pi_j(g_i, g_j) - \pi_i(g_i, g_j) = g_i - g_j$

Proof.

<u>First part of the proof: Proving lemma 0 (a)</u>

Let's consider player i makes an arbitrarily small contribution $\underline{g_i}$, and let further $g_j > \underline{g_i}$ be the case. Then, the material payoff of player i when contributing $\underline{g_i}$, given that the other player contributes g_j is given by:

$$\pi_i\left(\underline{g_i}, g_j\right) = 30 - \underline{g_i} + m \times \left(\underline{g_i} + g_j\right)$$

And the payoff of player j given $\underline{g_i}$ and g_j is equivalent to the following expression:

$$\pi_j\left(\underline{g_i}, g_j\right) = 30 - g_j + m \times \left(\underline{g_i} + g_j\right)$$

Subtracting the latter from the former, we get:

$$\pi_i\left(\underline{g_i}, g_j\right) - \pi_j\left(\underline{g_i}, g_j\right) = 30 - \underline{g_i} + m \times \left(\underline{g_i} + g_j\right) - \left\{30 - g_j + m \times \left(\underline{g_i} + g_j\right)\right\}$$

Expanding the curly brackets, we get:

$$\pi_i\left(\underline{g_i}, g_j\right) - \pi_j\left(\underline{g_i}, g_j\right) = 30 - \underline{g_i} + m \times \left(\underline{g_i} + g_j\right) - 30 + g_j - m \times \left(\underline{g_i} + g_j\right)$$

Simplifying, we get:

$$\pi_i\left(\underline{g_i}, g_j\right) - \pi_j\left(\underline{g_i}, g_j\right) = g_j - \underline{g_i}$$

Given that $\underline{g_i} < g_j$, it then follows that $g_j - \underline{g_i} > 0$. Hence,

$$\pi_i\left(\underline{g_i}, g_j\right) - \pi_j\left(\underline{g_i}, g_j\right) > 0$$

Bringing $\pi_j\left(\underline{g_i}, g_j\right)$ to the RHS, we get:

$$\pi_i\left(\underline{g_i}, g_j\right) > \pi_j\left(\underline{g_i}, g_j\right)$$

Which proves lemma 0 (a).

Second part of the proof: Proving lemma 0 (b)

Now, substituting $\underline{g_i}$ by g_i in the derivations above it is straightforward to see that

$$\pi_i(g_i, g_j) - \pi_j(g_i, g_j) = g_j - g_i$$

Additionally, multiplying both hand sides by -1 we can see that:

$$\pi_j(g_i, g_j) - \pi_i(g_i, g_j) = g_i - g_j$$

Which proves lemma 0 (b).

QED.

8.2.1.2. Homo Economicus preferences – Proof of proposition 1

Proposition 1. *If subject i maximizes the utility function $U_i^{HE}(g_i, g_j) = \pi_i(g_i, g_j)$, where $\pi_i(g_i, g_j)$ denotes the material payoff of person i for the strategy combination in which i contributes g_i and the other player g_j, subject i's optimal contributions will be $c_i^* = g_i = 0\ \forall g_j \in A_j$ (resp. $c_i^* = g_i = 30\ \forall g_j \in A_j$) in the SDG (resp. CIG).*

Proof.

To see this, note that $\frac{\partial U_i^{HE}(g_i, g_j)}{\partial g_i} = m - 1$, which is negative for any social dilemma (as $m < 1$) and positive for any CIG (as $m > 1$). Therefore, it follows that $c_i^* = g_i = 0\ \forall g_j \in A_j$ ($c_i^* = g_i = 30\ \forall g_j \in A_j$) is the solution to subject i's maximization problem in the SDG (CIG).

QED.

8.2.1.3. Inequality Aversion Preferences

8.2.1.3.1. Proof of proposition 2

Proposition 2. *If subject i maximizes the utility function* $U_i^{FS}\left(\pi_i(g_i, g_j), \pi_j(g_i, g_j)\right)$*, where* i *contributes* g_i *and the other player contributes* g_j*, then subject* i*'s contribution attitudes, denoted as* c_i^**, will be*

(i), in the Social Dilemma,

$$c_i^* = \begin{cases} g_i = 0\ \forall g_j \in A_j & iff\ \beta_i < 1 - \underline{m} \\ g_i = g_j \forall g_j \in A_j & iff\ \beta_i > 1 - \underline{m} \\ g_i \in [0, g_j]\ \forall g_j \in A_j & iff\ \beta_i = 1 - \underline{m} \end{cases}$$

(ii), in the Common Interest Game,

$$c_i^* = \begin{cases} g_i = 30\ \forall g_j \in A_j & iff\ \alpha_i < \overline{m} - 1 \\ g_i = g_j\ \forall\ g_j \in A_j & iff\ \alpha_i > \overline{m} - 1 \\ g_i \in [g_j, 30]\ \forall g_j \in A_j & iff\ \alpha_i = \overline{m} - 1 \end{cases}$$

Proof.

First part of the proof: proving (i)

Step 1: Recall necessary functions.

First, let's recall the utility function we use to measure inequality aversion preferences:

$$U_i^{\mathrm{FS}}(\pi_i, \pi_{\mathrm{j}}) := \pi_i - \alpha_i * Max\{\pi_{\mathrm{j}} - \pi_i, 0\} - \beta_i * Max\{\pi_i - \pi_{\mathrm{j}}, 0\}$$

Step 2: Calculate the utility function of player i for cases where $g_i < g_j$*.*

Let's assume that player i contributes less than player j. To keep the notation consistent throughout the text, let's denote such a contribution as $\underline{g_i}$. Then, the utility function of a Fehr-Schmidt player i will take the following form:

$$U_i^{FS}\left(\pi_i\left(\underline{g_i}, g_j\right), \pi_j\left(\underline{g_i}, g_j\right)\right) = \pi_i\left(\underline{g_i}, g_j\right) - \beta_i \times \left(\pi_i\left(\underline{g_i}, g_j\right) - \pi_j\left(\underline{g_i}, g_j\right)\right)$$

Substituting $\pi_i\left(\underline{g_i}, g_j\right) = 30 - \underline{g_i} + m \times \left(\underline{g_i} + g_j\right)$ in the first term of the RHS and using the results of lemma 0 (b) above to simplify the last term of the RHS, $U_i^{FS}\left(\pi_i, \pi_j\right)$ collapses to:

$$U_i^{FS}\left(\pi_i\left(\underline{g_i}, g_j\right), \pi_j\left(\underline{g_i}, g_j\right)\right) = 30 - \underline{g_i} + m \times \left(\underline{g_i} + g_j\right) - \beta_i \times \left(g_j - \underline{g_i}\right)$$

Step 3: Calculate the utility function of player i for cases where $g_i > g_j$.

Let's now consider the case where player i contributes more than player j, and let's denominate such a contribution as $\bar{g}_i > g_j$. Analogously to the previous step, substituting $\pi_i\left(\bar{g}_i, g_j\right) = 30 - \bar{g}_i + m \times \left(\bar{g}_i + g_j\right)$ in the first term of the RHS and using, again, the results from lemma 0 (b), we can rewrite the utility function as follows:

$$U_i^{FS}\left(\pi_i\left(\bar{g}_i, g_j\right), \pi_j\left(\bar{g}_i, g_j\right)\right) = 30 - \bar{g}_i + m \times \left(\bar{g}_i + g_j\right) - \alpha_i \times \left(\bar{g}_i - g_j\right)$$

Step 4: Write the utility function of player i for cases where $g_i = g_j$.

By lemma 0 (b), we know that $\pi_j\left(g_i, g_j\right) - \pi_i\left(g_i, g_j\right) = g_i - g_j$. Hence, whenever $g_i = g_j$, then $\pi_j\left(g_i, g_j\right) - \pi_i\left(g_i, g_j\right) = 0$. Substituting this into our utility function, we get:

$$U_i^{FS}\left(\pi_i\left(g_i, g_j\right), \pi_j\left(g_i, g_j\right)\right) = \pi_i\left(g_i, g_j\right) - \beta_i \times (0)$$

And, hence, $U_i^{FS}\left(\pi_i\left(g_i, g_j\right), \pi_j\left(g_i, g_j\right)\right) = U_i^{HE}\left(g_i, g_j\right) \; \forall g_i = g_j$.

Step 5: Write the utility function of player i for all possible cases of $g_i \gtreqless g_{-i}$.

Given the results of steps 2 to 4, we can then write the Fehr-Schmidt utility function more compactly as:

$$U_i^{FS}\left(\pi_i(g_i, g_j), \pi_j(g_i, g_j)\right) = \begin{cases} 30 - g_i + m \times (g_i + g_j) - \beta_i \times (g_j - g_i) \; if \; g_i < g_j \\ 30 - g_i + m \times (g_i + g_j) \; if \; g_i = g_j \\ 30 - g_i + m \times (g_i + g_j) - \alpha_i \times (g_i - g_j) \; if \; g_i > g_j \end{cases}$$

Step 6: Finding person i's first derivative with respect to g_i.

Taking the first derivative of the linear utility function with respect to g_i, we get

$$\frac{\partial U_i^{FS}\left(\pi_i(g_i, g_j), \pi_j(g_i, g_j)\right)}{\partial g_i} = \begin{cases} -1 + m + \beta_i \; if \; g_i < g_j \\ -1 + m \; if \; g_i = g_j \\ -1 + m - \alpha_i \; if \; g_i > g_j \end{cases}$$

Step 7: Impose in the previous derivative $m = \underline{m} < 1$.

Thus, for a generic value $\underline{m} \in \left(\frac{1}{n}, 1\right)$, the previous first derivative reads:

$$\frac{\partial U_i^{FS}\left(\pi_i(g_i, g_j), \pi_j(g_i, g_j)\right)}{\partial g_i} = \begin{cases} -1 + \underline{m} + \beta_i \; if \; g_i < g_j \\ -1 + \underline{m} \; if \; g_i = g_j \\ -1 + \underline{m} - \alpha_i \; if \; g_i > g_j \end{cases}$$

Step 8: Prove that $c_i^* = g_i > g_j$ is not optimal given all the potential values of α_i and $\underline{m}$.

As $\alpha_i \geq 0$ and $\underline{m} < 1$, then from the last step it follows that, if $g_i > g_j$, then

$$\frac{\partial U_i^{FS}\left(\pi_i(g_i, g_j), \pi_j(g_i, g_j)\right)}{\partial g_i} = -1 + \underline{m} - \alpha_i = -1 + (< 1) - (\geq 0) = (< 0) + (\leq 0) = (< 0).$$

It follows that the marginal utility will always be strictly negative for $g_i > g_j$, and, given the linearity of the utility function, person i's optimal contribution against g_j will never lie within the range defined by $g_i > g_j$.

Step 9: Give the range of values of β_i *for which the marginal utility is positive (resp. negative; resp. zero), given* $g_i < g_j$.

Turning to the case where $g_i < g_j$, we have three different outcomes:

When $g_i < g_j$, then

- $\frac{\partial U_i^{FS}}{\partial g_i} < 0\ iff\ \beta_i < 1 - \underline{m}$

- $\frac{\partial U_i^{FS}}{\partial g_i} > 0\ iff\ \beta_i > 1 - \underline{m}$

- $\frac{\partial U_i^{FS}}{\partial g_i} = 0\ iff\ \beta_i = 1 - m$

Step 10: Outline c_i^* *for an SDG (i.e., given* $\underline{m}$*) in lieu of the previous steps.*

Given steps 8 and 9, and the linearity of U_i^{FS}, i's best response against each potential g_j (that is, c_i^*) in the SDG will be given by:

$$c_i^* = \begin{cases} g_i = 0\ \forall g_j \in A_j & if\ \beta_i < 1 - \underline{m} \\ g_i \in [0, g_j]\ \forall g_j \in A_j & if\ \beta_i = 1 - \underline{m} \\ g_i = g_j\ \forall\ g_j \in A_j & if\ \beta_i > 1 - \underline{m} \end{cases}$$

This follows from three facts:

1. First, note that whenever $\beta_i < 1 - \underline{m}$, then $\left(\forall \langle g_i g_j \rangle \in A_i \times A_j\right), \frac{\partial U_i^{FS}\left(\pi_i(g_i, g_j), \pi_j(g_i, g_j)\right)}{\partial g_i} < 0$. Hence, $g_i = 0\ \forall g_j \in A_j$ will maximise person i's contribution against each possible g_j.
2. Second, note that, whenever $\beta_i = 1 - \underline{m}$, then $\left(\forall \langle g_i g_j \rangle \in A_i \times A_j\right), \frac{\partial U_i^{FS}\left(\pi_i(g_i, g_j), \pi_j(g_i, g_j)\right)}{\partial g_i} = 0\ iff\, g_i \in [0, g_j]$; implying that person i's utility for all $g_i \leq g_j$ will be the same; all being optimal contributions.

3. Third, note that, whenever $\beta_i < 1 - \underline{m}$, then $\left(\forall \langle g_i g_j \rangle \in A_i \times A_j\right), \frac{\partial U_i^{FS}\left(\pi_i(g_i,g_j),\pi_j(g_i,g_j)\right)}{\partial g_i} > 0\ iff\ g_i < g_j$ and $\frac{\partial U_i^{FS}\left(\pi_i(g_i,g_j),\pi_j(g_i,g_j)\right)}{\partial g_i} < 0\ iff\ g_i \geq g_j$. Hence, person i's utility will be maximised, in such cases, at $g_i = g_j$.

<u>*Second part of the proof: proving (ii)*</u>

Step 11: Impose in the derivative $m = \overline{m} > 1$.

For a generic value $\overline{m}$, the previous first derivative is equivalent to:

$$\frac{\partial U_i^{FS}\left(\pi_i(g_i, g_j), \pi_j(g_i, g_j)\right)}{\partial g_i} = \begin{cases} -1 + \overline{m} + \beta_i\ if\ g_i < g_j \\ -1 + \overline{m}\ if\ g_i = g_j \\ -1 + \overline{m} - \alpha_i\ if\ g_i > g_j \end{cases}$$

Step 12: Prove that $g_i < g_j$ is not optimal given all the potential values of β_i and $\overline{m}$.

As $\beta_i \geq 0$ and $\overline{m} > 1$, then from the derivate it follows that, if $g_i < g_j$, then

$$\frac{\partial U_i^{FS}\left(\pi_i(g_i, g_j), \pi_j(g_i, g_j)\right)}{\partial g_i} = -1 + \overline{m} + \beta_i = -1 + (> 1) + (\geq 0) = (> 0) + (\geq 0)$$
$$= (> 0)$$

It follows that the marginal utility will always be strictly positive for $g_i < g_j$; and, given the linearity of the utility function, person i's optimal contribution will never lie within the range defined by $g_i < g_j$.

Step 13: Give the range of values of α_i for which the marginal utility is positive (resp. negative; resp. zero) given $g_i > g_j$.

Turning to the case where $g_i > g_j$, we have three different outcomes:

- $\frac{\partial U_i^{FS}}{\partial g_i} < 0\ iff\ \alpha_i > \overline{m} - 1$

- $\frac{\partial U_i^{FS}}{\partial g_i} > 0 \; iff \; \alpha_i < \overline{m} - 1$

- $\frac{\partial U_i^{FS}}{\partial g_i} = 0 \; iff \; \alpha_i = \overline{m} - 1$

Step 14: Outline c_i^* for a CIG (i.e., given $\overline{m}$) in lieu of the previous steps

Given steps 12 and 13, and the linearity of U_i^{FS}, i's best response against g_j (that is, c_i^*) in the CIG will be given by:

$$c_i^* = \begin{cases} g_i = 30 \; \forall g_j \in A_j & iff \; \alpha_i < \overline{m} - 1 \\ g_i \in [g_j, 30] \; \forall g_j \in A_j & iff \; \alpha_i = \overline{m} - 1 \\ g_i = g_j \; \forall g_j \in A_j & iff \; \alpha_i > \overline{m} - 1 \end{cases}$$

This follows from three facts:

1. First, note that whenever $\alpha_i < \overline{m} - 1$, then $\left(\forall \langle g_i g_j \rangle \in A_i \times A_j\right), \frac{\partial U_i^{FS}\left(\pi_i(g_i, g_j), \pi_j(g_i, g_j)\right)}{\partial g_i} > 0$. Hence, $c_i^* = g_i = 30 \; \forall g_j \in A_j$ will maximise person i's contribution against each possible g_j.
2. Second, note that, whenever $\alpha_i = \overline{m} - 1$, then $\left(\forall \langle g_i g_j \rangle \in A_i \times A_j\right), \frac{\partial U_i^{FS}\left(\pi_i(g_i, g_j), \pi_j(g_i, g_j)\right)}{\partial g_i} = 0 \; iff \; g_i \in [g_j, 30]$, implying that person i's utility for all $g_i \geq g_j$ will be the same, all being optimal contributions.
3. Third, note that, whenever $\alpha_i > \overline{m} - 1$, then $\left(\forall \langle g_i g_j \rangle \in A_i \times A_j\right), \frac{\partial U_i^{FS}\left(\pi_i(g_i, g_j), \pi_j(g_i, g_j)\right)}{\partial g_i} < 0 \; iff \; g_i > g_j$ and $\frac{\partial U_i^{FS}\left(\pi_i(g_i, g_j), \pi_j(g_i, g_j)\right)}{\partial g_i} > 0 \; iff \; g_i < g_j$. Hence, person i's utility will be maximised, in such cases, at $g_i = g_j$.

QED.

8.2.1.3.2. Other results involving inequality aversion preferences

We use the results from proposition 2 to provide, in corollary 2.1, the precise contribution attitudes in the SDG and CIG that we use in chapter 4. Additionally, we provide another main result besides proposition 2. Namely, that for some joint values of $\underline{m}$ and $\overline{m}$ person i cannot be a perfect conditional cooperator (i.e., $g_i = g_j \; \forall \; g_j \in A_j$) in the SDG and an unconditional cooperator in the CIG (i.e., $g_i = 30 \; \forall g_j \in A_j$), as it would require a violation of the parameter restrictions of Fehr-Schmidt (i.e., it would require $\beta_i > \alpha_i$). Hence, inequality aversion cannot predict perfect conditional cooperation in the SDG and unconditional cooperation in the CIG. We summarise this second result in corollary2.2. Additionally, corollary 2.3 shows that, for the values of $\underline{m}$ and $\overline{m}$ used in the experiments of chapter 4, the inequality aversion model cannot predict conditional co-operation in the SDG and unconditional co-operation in the CIG.

Corollary 2.1. *If subject* i *maximizes the utility function* $U_i^{FS}\left(\pi_i(g_i, g_j), \pi_j(g_i, g_j)\right)$, *and* $\underline{m} = 0.6$ in the SDG and $\overline{m} = 1.2$ in the CIG, then

(a) has $\beta_i < 0.4$ *(resp.* $\beta_i = 0.4$*; resp.* $\beta_i > 0.4$*), then subject* i*'s cooperation attitude in the SDG will be* $c_i^* = g_i = 0\ \forall g_j \in A_j$ *(resp.* $c_i^* = g_i \in [0, g_j]\ \forall g_j \in A_j$*; resp.* $c_i^* = g_i = g_j\ \forall\ g_j \in A_j$*).*

(b) has $\alpha_i < 0.2$ *(resp.* $\alpha_i = 0.2$*; resp.* $\alpha_i > 0.2$*), then subject* i*'s cooperation attitude in the CIG will be* $c_i^* = g_i = 30\ \forall g_j \in A_j$ *(resp.* $c_i^* = g_i \in [g_j, 30]\ \forall g_j \in A_j$*; resp.* $c_i^* = g_i = g_j\ \forall\ g_j \in A_j$*).*

Proof.

Substituting $\underline{m} = 0.6$ and $\overline{m} = 1.2$ in the cooperation attitudes found in proposition 2, we get the two following expressions:

$$c_i^* = \begin{cases} g_i = 0\ \forall g_j \in A_j & if\ \beta_i < 1 - 0.6 \\ g_i \in [0, g_j]\ \forall g_j \in A_j & if\ \beta_i = 1 - 0.6 \\ g_i = g_j\ \forall\ g_j \in A_j & if\ \beta_i > 1 - 0.6 \end{cases}$$

$$c_i^* = \begin{cases} g_i = 30\ \forall g_j \in A_j & iff\ \alpha_i < 1.2 - 1 \\ g_i \in [g_j, 30]\ \forall g_j \in A_j & iff\ \alpha_i = 1.2 - 1 \\ g_i = g_j\ \forall g_j \in A_j & iff\ \alpha_i > 1.2 - 1 \end{cases}$$

Which, after simplifying, become:

$$c_i^* = \begin{cases} g_i = 0\ \forall g_j \in A_j & if\ \beta_i < 0.4 \\ g_i \in [0, g_j]\ \forall g_j \in A_j & if\ \beta_i = 0.4 \\ g_i = g_j\ \forall\ g_j \in A_j & if\ \beta_i > 0.4 \end{cases}$$

$$c_i^* = \begin{cases} g_i = 30\ \forall g_j \in A_j & iff\ \alpha_i < 0.2 \\ g_i \in [g_j, 30]\ \forall g_j \in A_j & iff\ \alpha_i = 0.2 \\ g_i = g_j\ \forall g_j \in A_j & iff\ \alpha_i > 0.2 \end{cases}$$

QED.

Corollary 2.2. *If subject* i *maximizes the utility function* $U_i^{FS}\left(\pi_i(g_i, g_j), \pi_j(g_i, g_j)\right)$*, where* i *contributes* g_i *and the other player contributes* g_j*, and further* $2 > \underline{m} + \overline{m}$ *holds true, then subject* i *will be a perfect conditional co-operator in the SD and an unconditional co-operator in the CIG iff* $\beta_i > \alpha_i$.

Proof.

Step 1: Provide the conditions for perfect conditional cooperation in the SDG and unconditional cooperation in the CIG.

Given proposition 2, Subject i will only be a perfect conditional cooperator (i.e., $g_i = g_j \; \forall \; g_j \in A_j$) in the SDG iff the following condition holds:

$$\beta_i > 1 - \underline{m}$$

Additionally, given proposition 2, Subject i will only be an unconditional cooperator (i.e., $g_i = 30 \; \forall \; g_j \in A_j$) in the CIG iff the following condition holds:

$$\alpha_i < \overline{m} - 1$$

Step 2: Establish the result by contradiction.

Assume $2 > \underline{m} + \overline{m}$, that subject i is a perfect conditional co-operator in the SD and an unconditional co-operator in the CIG, and that $\alpha_i > \beta_i$ holds true at the same time. Then, by using the two previous conditions and imposing $\alpha_i > \beta_i$, we would get:

$$\overline{m} - 1 > \mathrm{\alpha_i} > \beta_i > 1 - \underline{m}$$

From which it trivially follows that:

$$\overline{m} - 1 > 1 - \underline{m}$$

And, hence,

$$\overline{m} + \underline{m} > 2$$

Thus, if $2 > \underline{m} + \overline{m}$, subject i is a perfect conditional co-operator in the SD and an unconditional co-operator in the CIG, and $\alpha_i > \beta_i$ hold true at the same time, it must be that $2 > \underline{m} + \overline{m}$ and $2 < \underline{m} + \overline{m}$ hold true at the same time, which is a contradiction. Therefore, if subject i is a perfect conditional co-operator in the SD and an unconditional co-operator in the CIG, and it happens to be that $2 > \underline{m} + \overline{m}$, then $\alpha_i < \beta_i$ must be true.

QED.

8.2.1.4. Reciprocity preferences

8.2.1.4.1. Fixing some notation specific to sequential reciprocity

In the next pages we present the theoretical derivations for the reciprocity model of Dufwenberg and Kirchsteiger (2004). From now on, we use $(p', g_i = x; q', g_i \neq x)$ as a notation to describe the probabilities associated with contribution levels $g_i = x$ and $g_i \neq x$, which represent nothing but the first order beliefs. Hence, we use $(p', g_i = 0; q', g_i = 10; r', g_i = 20; 1 - p' - q' - r', g_i = 30)$ to refer to the probabilities associated to each of the possible contribution levels in our games. We denote the probabilities associated with second order beliefs as p'', q'', and so on. Additionally, in the contribution table task we assume that the contribution of the other person in each cell represents the first order belief with certainty of the responder. This is the case as, given the comment in Fischbacher et al (2001), the responses to each cell in the strategy method, given the incentive compatible mechanism used, can be seen as the responses of a second mover to each potential move of the first mover. And, given the belief updating mechanism in Dufwenberg and Kirchsteiger (2004), at each node the second mover updates his belief to reflect what has been played by the first mover, hence collapsing the first order belief to the strategy that led to the node being played.

As a reminder, below is the utility function of person i if person i were to follow Dufwenberg and Kirchsteiger's (2004) model of reciprocity:

$$\begin{aligned} U_i^{DK}(\pi_i, \pi_j) &= \pi_i\left(g_i(h), b_{ij}(h), c_{iji}(h)\right) \\ &= \pi_i\left(g_i(h), b_{ij}(h)\right) + Y_{i,j} \times \kappa_{ij}\left(g_i(h), b_{ij}(h)\right) \times \lambda_{iji}\left(b_{ij}(h), c_{iji}(h)\right) \end{aligned}$$

Where Y_{ij} is a parameter measuring the strength of reciprocal motivations, $\kappa_{ij}\left(g_{ij}(h), b_{ij}(h)\right)$ is a function measuring how kind is person i being with person j, $\lambda_{ij}\left(b_{ij}(h), c_{iji}(h)\right)$ is a function measuring how kind person i perceives person j is being towards him and $g_i(h)$, $b_{ij}(h)$ and $c_{ij}(h)$ are, respectively, the contribution, first- and second-order beliefs of person i at node h. Given that person i is a second mover, $b_{ij}(h)$ is updated to reflect the contribution level of the first mover, person j; being, hence, possible an alternative notation $b_{ij}(h) = g_j$.

C.2.2.1.4.2. Proof of proposition 3

Proposition 3. *If subject i maximizes the utility function* $U_i^{DK}\left(\pi_i(g_i, g_j), \pi_j(g_i, g_j)\right)$*, where i contributes* g_i*, the other player contributes* g_j*, and the other player moves first and subject i second, and where we denote* c_i^* *as subject i's optimal contribution, then subject i will*

(i), in the Social Dilemma,

(a) do $c_i^* = g_i = 0$ *against* $g_j \in \{0,10\}$ *regardless of* $Y_{i,j}$

(b) do $c_i^* = g_i = 0$ *against* $g_j \in \{20,30\}$ *iff* $Y_{i,j} < \frac{1-\underline{m}}{\underline{m}^2 \times (g_j - 15)}$

(c) do $c_i^* = g_i \in A_i$ *against* $g_j \in \{20,30\}$ *iff* $Y_{i,j} = \frac{1-\underline{m}}{\underline{m}^2 \times (g_j - 15)}$

(d) do $c_i^* = g_i = 30$ *against* $g_j \in \{20,30\}$ *iff* $Y_{i,j} > \frac{1-\underline{m}}{\underline{m}^2 \times (g_j - 15)}$

(ii), in the Common Interest Game,

(e) do $c_i^* = g_i = 30$ *against* $g_j = 30$ *regardless of* $Y_{i,j}$

(f) do $c_i^* = g_i = 0$ *against* $g_j \in \{0,10,20\}$*) iff* $Y_{i,j} > \frac{\overline{m}-1}{\overline{m}^2 \times (30 - g_j)}$

(g) do $c_i^* = g_i \in A_i$ *against* $g_j \in \{20,30\}$ *iff* $Y_{i,j} = \frac{\overline{m}-1}{\overline{m}^2 \times (30 - g_j)}$

(h) do $c_i^* = g_i = 30$ *against* $g_j \in \{0,10,20\}$*) iff* $Y_{i,j} < \frac{\overline{m}-1}{\overline{m}^2 \times (30 - g_j)}$

Proof.

The proof for this proposition is very long, so we start by summarising the approach we take before the reader engages with the reading of the proof. The first steps will involve computing the kindness and perceived kindness functions of person i for a generic level of the other person. The next steps will involve substituting those functional forms into $U_i^{DK}\left(\pi_i(g_i, g_j), \pi_j(g_i, g_{ji})\right)$ to get the utility function of person i in terms, only, of g_i and g_j. We, then, compute the first order derivative of $U_i^{DK}\left(\pi_i(g_i, g_j), \pi_j(g_i, g_{ji})\right)$ with respect to g_i to find the optimal contribution levels of g_i. This is done, as was the case with inequality aversion preferences, by assessing if the utility function is either increasing or decreasing in g_i at every level of g_j. We will carry out this process separately for the SDG and the CIG as the set of efficient strategies is different for both games, making the functional form of the kindness and perceived kindness functions to differ across games.

Step 1: find the kindness function (κ_{ij}) of subject i in the SDG.

At generic contribution levels g_j and g_i, we can write the kindness function as:

$$\kappa_{ij}\left(g_{ij}(h), b_{ij}(h)\right) = \pi_j\left(g_i, b_{ij}(h)\right) - \frac{\max \pi_j\left(g_i, b_{ij}(h)\right) + \min \pi_j\left(g_i, b_{ij}(h)\right)}{2}$$

Given that $\pi_j(g_i, g_j) = 30 - g_j + \underline{m} \times (g_i + g_j)$, taking the first derivative with respect to g_i, we get:

$$\frac{\partial \pi_j(g_i, g_j)}{\partial g_i} = \underline{m} > 0$$

Hence, the payoff of person j is increasing in g_i. This means that the payoff of person j will be maximised, given g_j, at the highest contribution level of person i and will be minimised at the lowest contribution level of person i. Those are, respectively, $g_i = 30$ and $g_i = 0$. Additionally, and given that person j is the first mover, then $b_{ij}(h) = g_j$. Hence, we can rewrite the kindness function as:

$$\kappa_{ij}\left(g_{ij}(h), b_{ij}(h) = g_j\right) = \pi_j(g_i, g_j) - \frac{\pi_j(g_i = 30, g_j) + \pi_j(g_i = 0, g_j)}{2}$$

Substituting $\pi_j(g_i, g_j)$ by the material payoff function outlined above, and g_i by 0 and 30 where appropriate, we get:

$$\begin{aligned}\kappa_{ij}&\left(g_{ij}(h), g_j\right)\\ &= 30 - g_j + \underline{m} \times (g_i + g_j)\\ &\quad - \frac{30 - g_j + \underline{m} \times (30 + g_j) + 30 - g_j + \underline{m} \times (g_j)}{2}\end{aligned}$$

Grouping the terms in the numerator, and taking $\underline{m}$ as a common factor in the numerator, we get:

$$\kappa_{ij}\left(g_{ij}(h), g_j\right) = 30 - g_j + \underline{m} \times \left(g_i + g_j\right) - \frac{60 - 2 \times g_j + \underline{m} \times \left(30 + 2 \times g_j\right)}{2}$$

Which can be rewritten as:

$$\kappa_{ij}\left(g_{ij}(h), g_j\right) = 30 - g_j + \underline{m} \times \left(g_i + g_j\right) - \left(30 - g_j + \underline{m} \times \left(15 + g_j\right)\right)$$

Expanding the expression $-\left(30 - g_j + \underline{m} \times \left(15 + g_j\right)\right)$, we get:

$$\kappa_{ij}\left(g_{ij}(h), g_j\right) = 30 - g_j + \underline{m} \times \left(g_i + g_j\right) - 30 + g_j - \underline{m} \times \left(15 + g_j\right)$$

Simplifying, we get:

$$\kappa_{ij}\left(g_{ij}(h), g_j\right) = \underline{m} \times \left(g_i + g_j\right) - \underline{m} \times \left(15 + g_j\right)$$

Using $\underline{m}$ as a common factor, we get:

$$\kappa_{ij}\left(g_{ij}(h), g_j\right) = \underline{m} \times \left(g_i + g_j - 15 - g_j\right)$$

And, finally, simplifying we get:

$$\kappa_{ij}\left(g_{ij}(h), g_j\right) = \underline{m} \times (g_i - 15)$$

Step 2: find the perceived kindness function (λ_{iji}) of subject i in the SDG.

To compute the perceived kindness function, let us denominate $c_{iji}(h) = (p'', g_i = 0; q'', g_i = 10; r'', g_i = 20; 1 - p'' - q'' - r'', g_i = 30)$ as the probability distribution of the second-order belief of player i. Unlike the first-order belief, the first mover did not know what player 2 was going to do when he or she decided to contribute g_j. Hence, we assume that the second mover believes that the first mover didn't know what the second mover was going to do when first mover chose g_j. The probability distribution $c_{iji}(h)$ over the second-order belief captures that uncertainty. We use such generic probability distribution to denote the belief that player i has about the belief of player j of player i's contribution when player j was making the decision of contributing g_j (contribution at the initial node). For compactness in the notation, we just write $c_{iji}(h)$ instead of writing $c_{iji}(h) = (p'', g_i = 0; q'', g_i = 10; r'', g_i = 20; 1 - p'' - q'' - r'', g_i = 30)$ in our definition of the perceived kindness function of person i. We can define the perceived kindness function of player i as:

$$\lambda_{iji}\left(b_{ij}(h), c_{iji}(h)\right) = \pi_i\left(b_{ij}(h), c_{iji}(h)\right) - \frac{\max \pi_i\left(b_{ij}(h), c_{iji}(h)\right) + \min \pi_i\left(b_{ij}(h), c_{iji}(h)\right)}{2}$$

As noted before, the payoff function of a given player is increasing in the contribution of the other player. Hence, person i's payoff will be maximised at $b_{ij}(h) = 30$ and minimised at $b_{ij}(h) = 0$. Hence, $\max \pi_i\left(b_{ij}(h), c_{iji}(h)\right) = \pi_i\left(30, c_{iji}(h)\right)$ and $\min \pi_i\left(b_{ij}(h), c_{iji}(h)\right) = \pi_i\left(0, c_{iji}(h)\right)$.

Given that $c_{iji}(h)$ is a probability distribution, then the payoff that person i beliefs that person j intends to give person i by contributing $b_{ij}(h) = g_j$ is an expected payoff of all the potential payoffs that person i could get for every action that person i makes weighted by the corresponding probability value in the probability distribution of $c_{iji}(h)$. In more intuitive terms, we can rewrite $\pi_i\left(g_j, c_{iji}(h)\right)$, $\pi_i\left(30, c_{iji}(h)\right)$ and $\pi_i\left(0, c_{iji}(h)\right)$ as follows:

$$\begin{aligned}\pi_i\left(g_j, c_{iji}(h)\right) &= p'' \times \pi_i\left(g_j, g_i = 0\right) + q'' \times \pi_i\left(g_j, g_i = 10\right) + r'' \times \pi_i\left(g_j, g_i = 20\right) \\ &\quad + (1 - p'' - q'' - r'') \times \pi_i\left(g_j, g_i = 30\right)\end{aligned}$$

$$\begin{aligned}\pi_i\left(30, c_{iji}(h)\right) &= p'' \times \pi_i(30, g_i = 0) + q'' \times \pi_i(30, g_i = 10) + r'' \times \pi_i(30, g_i = 20) \\ &\quad + (1 - p'' - q'' - r'') \times \pi_i(30, g_i = 30)\end{aligned}$$

$$\pi_i\left(0, c_{iji}(h)\right) = p'' \times \pi_i(0, g_i = 0) + q'' \times \pi_i(0, g_i = 10) + r'' \times \pi_i(0, g_i = 20) \\ + (1 - p'' - q'' - r'') \times \pi_i(0, g_i = 30)$$

Substituting each of the relevant elements of the RHS in each of the previous three equations by the corresponding material payoff function described earlier, we get:

$$\pi_i\left(g_j, c_{iji}(h)\right) \\ = p'' \times \left(30 - 0 + \underline{m} \times \left(g_j + 0\right)\right) + q'' \times \left(30 - 10 + \underline{m} \times \left(g_j + 10\right)\right) \\ + r'' \times \left(30 - 20 + \underline{m} \times \left(g_j + 20\right)\right) \\ + (1 - p'' - q'' - r'') \times \left(30 - 30 + \underline{m} \times \left(g_j + 30\right)\right)$$

$$\pi_i\left(30, c_{iji}(h)\right) \\ = p'' \times \left(30 - 0 + \underline{m} \times (0 + 30)\right) + q'' \times \left(30 - 10 + \underline{m} \times (30 + 10)\right) \\ + r'' \times \left(30 - 20 + \underline{m} \times (30 + 20)\right) \\ + (1 - p'' - q'' - r'') \times \left(30 - 30 + \underline{m} \times (30 + 30)\right)$$

$$\pi_i\left(0, c_{iji}(h)\right) = p'' \times \left(30 + \underline{m} \times (0 + 0)\right) + q'' \times \left(20 + \underline{m} \times (0 + 10)\right) \\ + r'' \times \left(30 - 20 + \underline{m} \times (0 + 20)\right) \\ + (1 - p'' - q'' - r'') \times \left(30 - 30 + \underline{m} \times (0 + 30)\right)$$

Which simplify to:

$$\pi_i\left(g_j, c_{iji}(h)\right) \\ = p'' \times \left(30 + \underline{m} \times \left(g_j\right)\right) + q'' \times \left(20 + \underline{m} \times \left(g_j + 10\right)\right) \\ + r'' \times \left(10 + \underline{m} \times \left(g_j + 20\right)\right) + (1 - p'' - q'' - r'') \times \left(\underline{m} \times \left(g_j + 30\right)\right)$$

$$\begin{aligned}\pi_i\left(30, c_{iji}(h)\right) \\ &= p'' \times \left(30 + \underline{m} \times 30\right) + q'' \times \left(20 + \underline{m} \times 40\right) + r'' \times \left(10 + \underline{m} \times 50\right) \\ &+ (1 - p'' - q'' - r'') \times \left(\underline{m} \times 60\right)\end{aligned}$$

$$\begin{aligned}\pi_i\left(0, c_{iji}(h)\right) = p'' \times (30) + q'' \times \left(20 + \underline{m} \times 10\right) + r'' \times \left(10 + \underline{m} \times 20\right) \\ + (1 - p'' - q'' - r'') \times \left(\underline{m} \times 30\right)\end{aligned}$$

Using the last two equations, and taking p'', q'', r'' and $1 - p'' - q'' - r''$ as common factors, we can express $\pi_i\left(30, c_{iji}(h)\right) + \pi_i\left(0, c_{iji}(h)\right)$ as:

$$\begin{aligned}\pi_i\left(30, c_{iji}(h)\right) + \pi_i\left(0, c_{iji}(h)\right) \\ &= p'' \times \left(30 + \underline{m} \times 30 + 30\right) + q'' \times \left(20 + \underline{m} \times 40 + 20 + \underline{m} \times 10\right) \\ &+ r'' \times \left(10 + \underline{m} \times 50 + 10 + \underline{m} \times 20\right) \\ &+ (1 - p'' - q'' - r'') \times \left(\underline{m} \times 60 + \underline{m} \times 30\right)\end{aligned}$$

Which can be simplified to:

$$\begin{aligned}\pi_i\left(30, c_{iji}(h)\right) + \pi_i\left(0, c_{iji}(h)\right) \\ &= p'' \times \left(60 + \underline{m} \times 30\right) + q'' \times \left(40 + \underline{m} \times 50\right) + r'' \times \left(20 + \underline{m} \times 70\right) \\ &+ (1 - p'' - q'' - r'') \times \left(\underline{m} \times 90\right)\end{aligned}$$

Hence, the second term of the perceived kindness function, $\frac{\pi_i\left(30, c_{iji}(h)\right) + \pi_i\left(0, c_{iji}(h)\right)}{2}$, can be written as:

$$\begin{aligned}\frac{\pi_i\left(30, c_{iji}(h)\right) + \pi_i\left(0, c_{iji}(h)\right)}{2} \\ &= p'' \times \left(30 + \underline{m} \times 15\right) + q'' \times \left(20 + \underline{m} \times 25\right) + r'' \times \left(10 + \underline{m} \times 35\right) \\ &+ (1 - p'' - q'' - r'') \times \underline{m} \times 45\end{aligned}$$

Now, using the expressions we found for $\pi_i\left(g_j, c_{iji}(h)\right)$ and $\frac{\pi_i\left(30, c_{iji}(h)\right)+\pi_i\left(0, c_{iji}(h)\right)}{2}$, and taking p'', q'', r'' and $1-p''-q''-r''$ as common factors, we can express the perceived kindness function as:

$$\begin{aligned}
&\lambda_{iji}\left(g_j, c_{iji}(h)\right)\\
&\qquad = p'' \times \left(30 + \underline{m} \times g_j - 30 - \underline{m} \times 15\right)\\
&\qquad + q'' \times \left(20 + \underline{m} \times \left(g_j + 10\right) - 20 - \underline{m} \times 25\right)\\
&\qquad + r'' \times \left(10 + \underline{m} \times \left(g_j + 20\right) - 10 - \underline{m} \times 35\right)\\
&\qquad + (1 - p'' - q'' - r'') \times \left(\underline{m} \times \left(g_j + 30\right) - \underline{m} \times 45\right)
\end{aligned}$$

By taking $\underline{m}$ as a common factor and simplifying, we get:

$$\begin{aligned}
&\lambda_{iji}\left(g_j, c_{iji}(h)\right)\\
&\qquad = p'' \times \left(\underline{m} \times \left(g_j - 15\right)\right) + q'' \times \left(\underline{m} \times \left(g_j + 10 - 25\right)\right)\\
&\qquad + r'' \times \left(\underline{m} \times \left(g_j + 20 - 35\right)\right)\\
&\qquad + (1 - p'' - q'' - r'') \times \left(\underline{m} \times \left(g_j + 30 - 45\right)\right)
\end{aligned}$$

Which can be further simplified to:

$$\begin{aligned}
&\lambda_{iji}\left(g_j, c_{iji}(h)\right)\\
&\qquad = p'' \times \left(\underline{m} \times \left(g_j - 15\right)\right) + q'' \times \left(\underline{m} \times \left(g_j - 15\right)\right)\\
&\qquad + r'' \times \left(\underline{m} \times \left(g_j - 15\right)\right) + (1 - p'' - q'' - r'') \times \left(\underline{m} \times \left(g_j - 15\right)\right)
\end{aligned}$$

Now, taking $\underline{m} \times \left(g_j - 15\right)$ as a common factor, we can rewrite the previous expression as:

$$\lambda_{iji}\left(g_j, c_{iji}(h)\right) = \underline{m} \times \left(g_j - 15\right) \times (p'' + q'' + r'' + 1 - p'' - q'' - r'')$$

Which can be further simplified to:

$$\lambda_{iji}\left(g_j, c_{iji}(h)\right) = \underline{m} \times \left(g_j - 15\right)$$

Step 3: Substitute the two expressions found in the reciprocity utility function.

Given the expressions of the kindness and perceived kindness function of person i, we can rewrite his or her utility as:

$$U_i^{DK}\left(\pi_i, \pi_j\right) = \pi_i\left(g_i(h), g_j, \kappa_{ij}, \lambda_{iji}\right) = \pi_i\left(g_i, g_j\right) + Y_{i,j} \times \underline{m} \times (g_i - 15) \times \underline{m} \times \left(g_j - 15\right)$$

Which, substituting $\pi_i\left(g_i, g_j\right)$ by the payoff function given g_i and g_j, we get:

$$U_i^{DK}\left(\pi_i, \pi_j\right) = 30 - g_i + \underline{m} \times \left(g_i + g_j\right) + Y_{i,j} \times \underline{m}^2 \times (g_i - 15) \times \left(g_j - 15\right)$$

Step 4: Compute the first order derivative of the utility function.

Taking the first derivative of the utility function with respect to the contribution of person i, we get:

$$\frac{\partial U_i^{DK}\left(\pi_i, \pi_j\right)}{\partial g_i} = -1 + \underline{m} + Y_{i,j} \times \underline{m}^2 \times \left(g_j - 15\right)$$

Step 5: Compute the sign of first order derivative of the utility function for $g_j \in \{0{,}10\}$.

When $g_j \in \{0{,}10\}$, then $g_j - 15 = (\leq 10) - 15 = (< 0)$. As $Y_{i,j} > 0$, and $\underline{m} < 1$, it, hence, follows that:

$$\frac{\partial U_i^{DK}\left(\pi_i, \pi_j\right)}{\partial g_i} = -1 + (< 1) + (\geq 0) \times \underline{m}^2 \times (< 0) = (< 0) + (< 0) = (< 0)$$

Hence,

$$\frac{\partial U_i^{DK}(\pi_i, \pi_j)}{\partial g_i} < 0$$

Which demonstrates that the utility function is decreasing over the whole domain of g_i for $g_j \in \{0,10\}$.

Step 6: Compute the optimal contribution of person i against $g_j \in \{0,10\}$.

Given that, for $g_j \in \{0,10\}$, the derivative of the utility function is negative over the whole domain of g_i, person i will maximise their utility by contributing nothing. That is,

$$(\forall\, Y_{i,j}), c_i^* = g_i = 0 \forall g_j \in \{0,10\}$$

Step 7: Compute the sign of first order derivative of the utility function for $g_j \in \{20,30\}$ in terms of $Y_{i,j}$.

The marginal utility becomes negative iff:

$$-1 + \underline{m} + Y_{i,j} \times \underline{m}^2 \times (g_j - 15) < 0$$

Isolating $Y_{i,j}$ if the LHS, we get:

$$Y_{i,j} \times \underline{m}^2 \times (g_j - 15) < 1 - \underline{m}$$

Dividing both sides by $\underline{m}^2 \times (g_j - 15)$, we get:

$$Y_{i,j} < \frac{1 - \underline{m}}{\underline{m}^2 \times (g_j - 15)}\ iff\ \frac{\partial U_i^{DK}(\pi_i, \pi_j)}{\partial g_i} < 0$$

For $g_j \in \{20{,}30\}$, whenever $Y_{i,j}$ is lower than the threshold value found above, the marginal utility with respect to g_i will be negative. In contrast, whenever the marginal utility is positive, we get the following condition:

$$Y_{i,j} > \frac{1 - \underline{m}}{\underline{m}^2 \times (g_j - 15)} \; iff \; \frac{\partial U_i^{DK}(\pi_i, \pi_j)}{\partial g_i} < 0$$

And whenever the marginal utility is exactly 0, it then follows that:

$$Y_{i,j} = \frac{1 - \underline{m}}{\underline{m}^2 \times (g_j - 15)} \; iff \; \frac{\partial U_i^{DK}(\pi_i, \pi_j)}{\partial g_i} = 0$$

Step 8: Compute the optimal contribution of person i *against* $g_j \in \{20{,}30\}$ *for all possible values of* $Y_{i,j}$*.*

Given the inequalities found in the previous step, the best responses against $g_j \in \{20{,}30\}$ can be summarised as:

$$c_i^* = \begin{cases} g_i = 0 \; \forall g_j \in \{20{,}30\} \; iff \; Y_{i,j} < \dfrac{1 - \underline{m}}{\underline{m}^2 \times (g_\mathrm{j} - 15)} \\ g_i \in A_i \; \forall g_j \in \{20{,}30\} \; iff \; Y_{i,j} = \dfrac{1 - \underline{m}}{\underline{m}^2 \times (g_\mathrm{j} - 15)} \\ g_i = 30 \; \forall g_j \in \{20{,}30\} \; iff \; Y_{i,j} > \dfrac{1 - \underline{m}}{\underline{m}^2 \times (g_\mathrm{j} - 15)} \end{cases}$$

Where the previous results hold given the linearity of the utility function $U_i^{DK}(\pi_i, \pi_j)$. That is, whenever the derivative is decreasing in the whole domain of g_i, as it is the case of the first of the two equations, then the best answer is to free ride; and whenever the derivative is increasing in the whole domain of g_i, as is the case of the second of the two equations, the best answer is to fully contribute. Whenever the derivative is equal to zero, any contribution gives the same utility and hence all are optimal choices. The sign of the derivative is determined by the reciprocity parameter $Y_{i,j}$.

Step 9: show that only full contribution (i.e., $g_i = 30$*) is an efficient strategy in the CIG.*

Unlike in the SDG, now only full contribution is an efficient strategy in a common interest game. This is the case as, for each and every of the contributions of the first mover player j – that is, for each of the possible histories of play before player i gets to play –, full contribution by player i gives no lower material payoff to any player and a higher material payoff to all players. As Player i's contribution decision is the only subsequent play for each and every contribution of player j, then by Dufwenberg and Kirchsteiger's (2004, pp. 276) definition of the set of efficient strategies, it follows that full contribution is the only strategy within the set of efficient strategies of player i, $E_i = \{g_i = 30\}$.

To see why $g_i = 30$ gives no lower material payoff to any of the players, notice that, in a common interest game,$\overline{m} \in (1, \infty)$. Hence, start by assuming that $g_i = 30$ implies

$$\pi_i(30, g_j) > \pi_i\left(\underline{g_i}, g_j\right)$$

Substituting the material payoff function by its functional form yields:

$$\overline{m} \times (30 + g_j) > 30 - \underline{g_i} + \overline{m} \times \left(\underline{g_i} + g_j\right)$$

Where $\underline{g_i} < 30$ is an arbitrarily small contribution of player i. Bringing $\overline{m}$ to the LHS, and taking $\overline{m}$ as a common factor, we get:

$$\overline{m} \times \left(30 + g_j - \underline{g_i} - g_j\right) > 30 - \underline{g_i}$$

Simplifying the parenthesis in the LHS, we get:

$$\overline{m} \times \left(30 - \underline{g_i}\right) > 30 - \underline{g_i}$$

Dividing both hand sides by $\left(30 - \underline{g_i}\right)$, we get:

$$\overline{m} > 1$$

Which is exactly the condition that will always hold in common interest games, thereby discharging the initial assumption. Hence, it follows that $g_i = 30$ gives the highest material payoff to player i.

Now, consider the payoff function of player j:

$$\pi_j(g_i, g_j) = 30 - g_j + \overline{m} \times (g_i + g_j)$$

The derivative of the function with respect to g_i is given by:

$$\frac{\partial \pi_j(g_i, g_j)}{\partial g_j} = \overline{m}$$

As $\overline{m} \in (1, \infty)$ in common interest games, it follows that $\frac{\partial \pi_j(g_i, g_j)}{\partial g_j} = \overline{m} > 0$. As the payoff function is linear in the contribution of player i and it is also increasing in it, it follows that $g_i = 30$ is the contribution of player i that will maximise the payoff of player j.

Hence, it follows that there doesn't exist another g_i that gives a higher payoff to any of the players, thereby proving why $g_i = 30$ is the only efficient strategy in common interest games.

Step 10: Outline the implications of a reduced set of efficient strategies in the kindness function (κ_{ij}) and the perceived kindness function (λ_{iji}) of subject i in the CIG.

This has important implications when computing the equitable payoff in both the kindness and perceived kindness functions, as the minimum payoff that can be given to any player is evaluated within the strategies that are efficient. Hence,

$$\min \pi_j(g_i, b_{ij}(h) = g_j) | g_i \in E_i = \max \pi_j(g_i, b_{ij}(h) = g_j) | g_i \in A_i$$
$$= \pi_j(g_i = 30, b_{ij}(h) = g_j)$$

and

$\max \pi_i \left(b_{ij}(h) = g_j, c_{iji}(h)\right) | g_j \in A_j = \min \pi_i \left(b_{ij}(h) = g_j, c_{iji}(h)\right) | g_j \in E_j =$
$\pi_i \left(b_{ij}(h) = 30, c_{iji}(h)\right)$.

The implication for the kindness and perceived kindness functions is that they can be defined as:

$$\kappa_{ij}\left(g_i, b_{ij}(h)\right) = \pi_j(g_i, g_j) - \frac{2 \times \pi_j(30, g_j)}{2}$$

$$\lambda_{iji}\left(b_{ij}(h), c_{iji}(h)\right) = \pi_i\left(g_j, c_{iji}(h)\right) - \frac{2 \times \pi_i\left(30, c_{iji}(h)\right)}{2}$$

Which can be simplified to:

$$\kappa_{ij}\left(g_i, b_{ij}(h)\right) = \pi_j(g_i, g_j) - \pi_j(30, g_j)$$

$$\lambda_{iji}\left(b_{ij}(h), c_{iji}(h)\right) = \pi_i\left(g_j, c_{iji}(h)\right) - \pi_i\left(30, c_{iji}(h)\right)$$

Step 11: find the kindness function (κ_{ij}) of subject i in the CIG.

At generic contribution levels g_j and g_i, then $b_{ij}(h) = g_j$. Hence, we can write the kindness function as:

$$\kappa_{ij}\left(g_{ij}(h), b_{ij}(h)\right) = \pi_j(g_i, g_j) - \pi_j(30, g_j)$$

Substituting $\pi_j(g_i, g_j)$ by the payoff function outlined above, we get:

$$\kappa_{ij}\left(g_{ij}(h), b_{ij}(h)\right) = 30 - g_j + \overline{m} \times (g_i + g_j) - 30 + g_j - \overline{m} \times (30 + g_j)$$

Simplifying, we get:

$$\kappa_{ij}\left(g_{ij}(h), b_{ij}(h)\right) = \overline{m} \times (g_i + g_j) - \overline{m} \times (30 + g_j)$$

Using $\overline{m}$ as a common factor, we get:

$$\kappa_{ij}\left(g_{ij}(h), b_{ij}(h)\right) = \overline{m} \times \left(g_i + g_j - 30 - g_j\right)$$

And, finally, simplifying we get:

$$\kappa_{ij}\left(g_{ij}(h), b_{ij}(h)\right) = \overline{m} \times (g_i - 30)$$

Step 12: find the perceived kindness function (λ_{iji}) of subject i in the CIG.

We can define the perceived kindness function of player i as:

$$\lambda_{iji}\left(b_{ij}(h), c_{iji}(h)\right) = \pi_i\left(g_j, c_{iji}(h)\right) - \pi_i\left(30, c_{iji}(h)\right)$$

Given that $c_{iji}(h)$ is the probability distribution described earlier, we can rewrite $\pi_i\left(g_j, c_{iji}(h)\right)$ and $\pi_i\left(30, c_{iji}(h)\right)$ as follows:

$$\begin{aligned}\pi_i\left(g_j, c_{iji}(h)\right) \\ &= p'' \times \pi_i\left(g_j, g_i = 0\right) + q'' \times \pi_i\left(g_j, g_i = 10\right) + r'' \times \pi_i\left(g_j, g_i = 20\right) \\ &+ (1 - p'' - q'' - r'') \times \pi_i\left(g_j, g_i = 30\right)\end{aligned}$$

$$\begin{aligned}\pi_i\left(30, c_{iji}(h)\right) \\ &= p'' \times \pi_i(30, g_i = 0) + q'' \times \pi_i(30, g_i = 10) + r'' \times \pi_i(30, g_i = 20) \\ &+ (1 - p'' - q'' - r'') \times \pi_i(30, g_i = 30)\end{aligned}$$

Substituting each of the elements of the RHS in each of the previous three equations by the corresponding payoff function described earlier, we get:

$$\pi_i\left(g_j, c_{iji}(h)\right)$$
$$= p'' \times \left(30 - 0 + \overline{m} \times \left(g_j + 0\right)\right) + q'' \times \left(30 - 10 + \overline{m} \times \left(g_j + 10\right)\right)$$
$$+ r'' \times \left(30 - 20 + \overline{m} \times \left(g_j + 20\right)\right)$$
$$+ (1 - p'' - q'' - r'') \times \left(30 - 30 + \overline{m} \times \left(g_j + 30\right)\right)$$

$$\pi_i\left(30, c_{iji}(h)\right)$$
$$= p'' \times \left(30 - 0 + \overline{m} \times (0 + 30)\right) + q'' \times \left(30 - 10 + \overline{m} \times (30 + 10)\right)$$
$$+ r'' \times \left(30 - 20 + \overline{m} \times (30 + 20)\right)$$
$$+ (1 - p'' - q'' - r'') \times \left(30 - 30 + \overline{m} \times (30 + 30)\right)$$

Which simplify to:

$$\pi_i\left(g_j, c_{iji}(h)\right)$$
$$= p'' \times \left(30 + \overline{m} \times \left(g_j\right)\right) + q'' \times \left(20 + \overline{m} \times \left(g_j + 10\right)\right)$$
$$+ r'' \times \left(10 + \overline{m} \times \left(g_j + 20\right)\right) + (1 - p'' - q'' - r'') \times \left(\overline{m} \times \left(g_j + 30\right)\right)$$

$$\pi_i\left(30, c_{iji}(h)\right)$$
$$= p'' \times (30 + \overline{m} \times 30) + q'' \times (20 + \overline{m} \times 40) + r'' \times (10 + M\overline{m} \times 50)$$
$$+ (1 - p'' - q'' - r'') \times (\overline{m} \times 60)$$

Now, using the expressions we found for $\pi_i\left(g_j, c_{iji}(h)\right)$ and $\pi_i\left(30, c_{iji}(h)\right)$, and taking p'', q'', r'' and $1 - p'' - q'' - r''$ as common factors, we can express the perceived kindness function as:

$$\begin{aligned}
&\lambda_{iji}\left(b_{ij}(h), c_{iji}(h)\right) \\
&\quad = p'' \times \left(30 + \overline{m} \times g_j - 30 - \overline{m} \times 30\right) \\
&\quad + q'' \times \left(20 + \overline{m} \times \left(g_j + 10\right) - 20 - \overline{m} \times 40\right) \\
&\quad + r'' \times \left(10 + \overline{m} \times \left(g_j + 20\right) - 10 - \overline{m} \times 50\right) \\
&\quad + (1 - p'' - q'' - r'') \times \left(\overline{m} \times \left(g_j + 30\right) - \overline{m} \times 60\right)
\end{aligned}$$

By taking $\overline{m}$ as a common factor and simplifying, we get:

$$\begin{aligned}
&\lambda_{iji}\left(b_{ij}(h), c_{iji}(h)\right) \\
&\quad = p'' \times \left(\overline{m} \times \left(g_j - 30\right)\right) + q'' \times \left(\overline{m} \times \left(g_j + 10 - 40\right)\right) \\
&\quad + r'' \times \left(\overline{m} \times \left(g_j + 20 - 50\right)\right) \\
&\quad + (1 - p'' - q'' - r'') \times \left(\overline{m} \times \left(g_j + 30 - 60\right)\right)
\end{aligned}$$

Which can be further simplified to:

$$\begin{aligned}
&\lambda_{iji}\left(b_{ij}(h), c_{iji}(h)\right) \\
&\quad = p'' \times \left(\overline{m} \times \left(g_j - 30\right)\right) + q'' \times \left(\overline{m} \times \left(g_j - 30\right)\right) \\
&\quad + r'' \times \left(\overline{m} \times \left(g_j - 30\right)\right) + (1 - p'' - q'' - r'') \times \left(\overline{m} \times \left(g_j - 30\right)\right)
\end{aligned}$$

Now, taking $\overline{m} \times \left(g_j - 30\right)$ as a common factor, we can rewrite the previous expression as:

$$\lambda_{iji}\left(b_{ij}(h) = g_j, c_{iji}(h)\right) = \overline{m} \times \left(g_j - 30\right) \times (p'' + q'' + r'' + 1 - p'' - q'' - r'')$$

Which can be further simplified to:

$$\lambda_{iji}\left(b_{ij}(h), c_{iji}(h)\right) = \overline{m} \times \left(g_j - 30\right)$$

Step 13: Substitute the two expressions found in the reciprocity utility function.

Given the expressions of the kindness and perceived kindness function of person i, we can rewrite his or her utility as:

$$\begin{aligned} U_i^{DK}(\pi_i, \pi_j) &= \pi_i\left(g_i, g_j, b_{ij}(h), c_{iji}(h)\right) \\ &= \pi_i\left(g_i, b_{ij}(h)\right) + Y_{i,j} \times \overline{m} \times (g_i - 30) \times \overline{m} \times \left(g_j - 30\right) \end{aligned}$$

Which, substituting $\pi_i\left(g_i, b_{ij}(h)\right)$ by the material payoff function given g_i and g_j, for a generic first-order belief of $g_j$we get:

$$U_i^{DK}(\pi_i, \pi_j) = 30 - g_i + \overline{m} \times \left(g_i + g_j\right) + Y_{i,j} \times \overline{m}^2 \times (g_i - 30) \times \left(g_j - 30\right)$$

Step 13: find the first order derivative of the utility function with respect to g_i.

Taking the first derivative of the utility function with respect to the contribution of person i, we get:

$$\frac{\partial U_i^{DK}(\pi_i, \pi_j)}{\partial g_i} = -1 + \overline{m} + Y_{i,j} \times \overline{m}^2 \times \left(g_j - 30\right)$$

Step 14: find the optimal contribution for person i against $g_j = 30$.

Note that, whenever $g_j = 30$, then $g_j - 30 = 0$. Hence, the reciprocal term collapses to 0 regardless of the value of $Y_{i,j}$. Hence, when $g_j = 30$ the marginal utility of own contribution is given by:

$$\left.\frac{\partial U_i^{DK}(\pi_i, \pi_j)}{\partial g_i}\right|_{g_j=30} = -1 + \overline{m}$$

As $\overline{m} > 1$ it follows that the marginal utility of own contribution when $g_j = 30$ will always be positive:

$$\left.\frac{\partial U_i^{DK}(\pi_i, \pi_j)}{\partial g_i}\right|_{g_j=30} = -1 + (> 1) = (> 0)$$

This implies that the best response against $g_j = 30$, given the linearity of the utility function with respect to own contribution, will be

$$(\forall\, Y_{i,j}), c_i^* = g_i = 30\; if\; g_j = 30$$

Step 15: find the optimal contribution for person i *against* $g_j \in \{0,10,20\}$.

Turning to the remaining cases, that is $g_j \in \{0,10,20\}$, we need to find for which values of $Y_{i,j}$ the marginal utility becomes negative. Recalling the marginal utility of g_i, we can capture that case with the following inequality:

$$-1 + \overline{m} + Y_{i,j} \times \overline{m}^2 \times (g_j - 30) < 0$$

Isolating $Y_{i,j}$ in the RHS, we get:

$$Y_{i,j} \times \overline{m}^2 \times (30 - g_j) > \overline{m} - 1$$

Dividing both sides by $\overline{m}^2 \times (30 - g_j)$, we get:

$$Y_{i,j} > \frac{\overline{m} - 1}{\overline{m}^2 \times (30 - g_j)}$$

For $g_j \in \{0,10,20\}$, then, we can capture person i's best responses as:

$$c_i^* = \begin{cases} g_i = 0\ \forall g_j \in \{0,10,20\}\ iff\ Y_{i,j} > \dfrac{\overline{m} - 1}{\overline{m}^2 \times (30 - g_j)} \\ g_i \in A_i\ \forall g_j \in \{0,10,20\}\ iff\ Y_{i,j} = \dfrac{\overline{m} - 1}{\overline{m}^2 \times (30 - g_j)} \\ g_i = 30\ \forall g_j \in \{0,10,20\}\ iff\ Y_{i,j} > \dfrac{\overline{m} - 1}{\overline{m}^2 \times (30 - g_j)} \end{cases}$$

QED.

8.2.1.4.3. Other results involving reciprocity preferences

We use the results from proposition 3 to provide, in corollary 3.1, the precise contribution attitudes in the SDG and CIG that we use in chapter 4. Additionally, we provide another main result besides proposition 3. Namely, that for some joint values of $\underline{m}$ and $\overline{m}$ person i cannot be a conditional cooperator in the SDG without being a conditional cooperator in the CIG. Hence, for such values of $\underline{m}$ and $\overline{m}$ preferences for reciprocity cannot predict conditional cooperation in the SDG and unconditional cooperation in the CIG. We summarise this statement in corollary3.2. Additionally, corollary 3.3 shows that, for the values of $\underline{m}$ and $\overline{m}$ used in the experiments of chapter 4, the result from corollary 3.2 holds true in our data. That is, preferences for reciprocity cannot rationalise conditional cooperation in the SDG and unconditional cooperation in the CIG.

Corollary 2.1. *If subject i maximizes the utility function* $U_i^{DK}\left(\pi_i(g_i,g_j),\pi_j(g_i,g_j)\right)$*, where i contributes* g_i*, the other player contributes* g_j*, and the other player moves first and subject i second, and where we denote* c_i^* *as subject i's optimal contribution schedule, then subject i will*

(i), in the Social Dilemma,

(a) do $c_i^* = g_i = 0$ *against* $g_j \in \{0{,}10\}$ *regardless of* $Y_{i,j}$

(b) do $c_i^* = g_i = 0$ *against* $g_j = 20$ *iff* $Y_{i,j} < \frac{0.4}{1.8}$

(c) do $c_i^* = g_i = 30$ *against* $g_j = 20$ *iff* $Y_{i,j} > \frac{0.4}{1.8}$

(d) do $c_i^* = g_i = 0$ *against* $g_j = 30$ *iff* $Y_{i,j} < \frac{0.4}{5.4}$

(e) do $c_i^* = g_i = 30$ *against* $g_j = 30$ *iff* $Y_{i,j} > \frac{0.4}{5.4}$

(ii), in the Common Interest Game,

(f) do $c_i^* = g_i = 30$ *against* $g_j = 30$ *regardless of* $Y_{i,j}$

(g) do $c_i^* = g_i = 0$ *against* $g_j = 0$*) iff* $Y_{i,j} > \frac{0.2}{1.2^2\times(30)}$

(h) do $c_i^* = g_i = 30$ *against* $g_j = 0$*) iff* $Y_{i,j} < \frac{0.2}{1.2^2\times(30)}$

(i) do $c_i^* = g_i = 0$ *against* $g_j = 10$*) iff* $Y_{i,j} > \frac{0.2}{1.2^2\times(20)}$

(j) do $c_i^* = g_i = 30$ *against* $g_j = 10$*) iff* $Y_{i,j} < \frac{0.2}{1.2^2\times(20)}$

(k) do $c_i^* = g_i = 0$ *against* $g_j = 20$*) iff* $Y_{i,j} > \frac{0.2}{1.2^2\times(10)}$

(l) do $c_i^* = g_i = 30$ *against* $g_j = 20$*) iff* $Y_{i,j} < \frac{0.2}{1.2^2\times(10)}$

Proof.

Given the contribution attitudes found in proposition 3, (a) and (f) follow without further demonstration. Substituting $\underline{m} = 0.6$ in the cooperation attitudes found in proposition 3, we get the following expressions for the SDG:

$c_i^* = g_i = 0$ against $g_j \in \{20{,}30\}$ iff $Y_{i,j} < \frac{1-0.6}{0.6^2\times(g_j-15)}$

$c_i^* = g_i = 30$ against $g_j \in \{20{,}30\}$ iff $Y_{i,j} > \frac{1-0.6}{0.36\times(g_j-15)}$

Substituting g_j explicitly in the inequalities, we get:

$c_i^* = g_i = 0$ against $g_j = 20$ iff $Y_{i,j} < \frac{1-0.6}{0.36\times(5)}$

$c_i^* = g_i = 30$ against $g_j = 20$ iff $Y_{i,j} > \frac{1-0.6}{0.36\times(5)}$

$c_i^* = g_i = 0$ against $g_j = 30$ iff $Y_{i,j} < \frac{1-0.6}{0.36\times(15)}$

$c_i^* = g_i = 30$ against $g_j = 30$ iff $Y_{i,j} > \frac{1-0.6}{0.36\times(15)}$

And, simplifying, we get:

$c_i^* = g_i = 0$ against $g_j = 20$ iff $Y_{i,j} < \frac{0.4}{1.8}$

$c_i^* = g_i = 30$ against $g_j = 20$ iff $Y_{i,j} > \frac{0.4}{1.8}$

$c_i^* = g_i = 0$ against $g_j = 30$ iff $Y_{i,j} < \frac{0,4}{5.4}$

$c_i^* = g_i = 30$ against $g_j = 30$ iff $Y_{i,j} > \frac{0.4}{5.4}$

Which proves (b), (c), (d), and (e). Additionally, substituting $\overline{m} = 1.2$ in the cooperation attitudes found in proposition 3, we get the following expressions for the CIG:

$c_i^* = g_i = 0$ against $g_j \in \{0,10,20\}$) iff $Y_{i,j} > \frac{1.2-1}{1.2^2\times(30-g_j)}$

$c_i^* = g_i = 30$ against $g_j \in \{0,10,20\}$) iff $Y_{i,j} < \frac{1.2-1}{1.2^2\times(30-g_j)}$

Substituting g_j explicitly in the inequalities, we get:

$c_i^* = g_i = 0$ against $g_j = 0$ iff $Y_{i,j} > \frac{1.2-1}{1.2^2\times(30)}$

$c_i^* = g_i = 30$ against $g_j = 0$ iff $Y_{i,j} < \frac{1.2-1}{1.2^2\times(30)}$

$c_i^* = g_i = 0$ against $g_j = 10$ iff $Y_{i,j} > \frac{1.2-1}{1.2^2\times(20)}$

$c_i^* = g_i = 30$ against $g_j = 10$ iff $Y_{i,j} < \frac{1.2-1}{1.2^2\times(20)}$

$c_i^* = g_i = 0$ against $g_j = 20$ iff $Y_{i,j} > \frac{1.2-1}{1.2^2\times(10)}$

$c_i^* = g_i = 30$ against $g_j = 20$ iff $Y_{i,j} < \frac{1.2-1}{1.2^2 \times (10)}$

And, simplifying, we get:

$c_i^* = g_i = 0$ against $g_j = 0$ iff $Y_{i,j} > \frac{0.2}{1.2^2 \times (30)}$

$c_i^* = g_i = 30$ against $g_j = 0$ iff $Y_{i,j} < \frac{0.2}{1.2^2 \times (30)}$

$c_i^* = g_i = 0$ against $g_j = 10$ iff $Y_{i,j} > \frac{0.2}{1.2^2 \times (20)}$

$c_i^* = g_i = 30$ against $g_j = 10$ iff $Y_{i,j} < \frac{0.2}{1.2^2 \times (20)}$

$c_i^* = g_i = 0$ against $g_j = 20$ iff $Y_{i,j} > \frac{0.2}{1.2^2 \times (10)}$

$c_i^* = g_i = 30$ against $g_j = 20$ iff $Y_{i,j} < \frac{0.2}{1.2^2 \times (10)}$

Which proves (g), (h), (i), (j), (k), and (l).

QED.

Corollary 2.2. *If subject i maximizes the utility function* $U_i^{DK}\left(\pi_i(g_i, g_j), \pi_j(g_i, g_j)\right)$*, where i contributes* g_i *and the other player contributes* g_j*, then if*

(i) *person i plays the weakest form of conditional cooperation possible in the SDG, and*

(ii) *it comes to pass that* $\frac{1-\underline{m}}{\underline{m}^2 \times (15)} > \frac{\overline{m}-1}{30 \times \overline{m}^2}$*,*

then subject i must play at least the weakest form of conditional cooperation in the CIG.

Proof.

Given proposition 3, the weakest conditional cooperation pattern predicted by reciprocity in the SDG entails subject i to fully contribute against full contribution and free ride otherwise. More formally, it entails subject i to play $c_i^* = g_i = 0$ against $g_j = \{0,10,20\}$ and $c_i^* = g_i = 30$ against $g_j = 30$ in the SDG. Also, the weakest form of conditional cooperation in the CIG entails free riding against free riding and full contribution otherwise. More formally, it entails subject i to play $c_i^* = g_i = 0$ against $g_j = 0$ and $c_i^* = g_i = 30$ against $g_j \in \{10,20,30\}$ in the CIG.

Given proposition 3, the referred pattern of cooperation attitude in the SDG holds iff:

$$Y_{i,j} > \frac{1-\underline{m}}{\underline{m}^2 \times (15)}$$

Then, given that $Y_{i,j} > \frac{1-\underline{m}}{\underline{m}^2 \times (15)}$ and that condition (ii) entails $\frac{1-\underline{m}}{\underline{m}^2 \times (15)} > \frac{\overline{m}-1}{30 \times \overline{m}^2}$, it naturally follows that:

$$Y_{i,j} > \frac{1-\underline{m}}{\underline{m}^2 \times (15)} > \frac{\overline{m}-1}{30 \times \overline{m}^2} \rightarrow Y_{i,j} > \frac{\overline{m}-1}{30 \times \overline{m}^2}$$

Recall that, given proposition 3, it follows that playing $c_i^* = g_i = 0$ against $g_j = 0$ and $c_i^* = g_i = 30$ against $g_j \in \{10,20,30\}$ in the CIG reveals the following inequality regarding $Y_{i,j}$:

$$Y_{ij} > \frac{\overline{m} - 1}{\overline{m}^2 \times \left(30 - g_j\right)}$$

Hence, it follows that for a subject maximizing U_i^{DK}, playing the weakest form of conditional cooperation in the SDG implies at least some conditional cooperation in the CIG.

QED.

Corollary 2.3. *Given* $\underline{m} = 0.6$ *and* $\overline{m} = 1.2$, *then the weakest form of conditional cooperation in the SDG implies at least a form of conditional cooperation in the CIG.*

Proof.

Recall from corollary 2.2 that, given the weakest form of conditional cooperation, if $\frac{1-\underline{m}}{\underline{m}^2\times(15)} > \frac{\overline{m}-1}{30\times\overline{m}^2}$ then reciprocity would predict conditional cooperation in the CIG. Substituting $\underline{m} = 0.6$ and $\overline{m} = 1.2$ in that condition, we get:

$$\frac{1-0.6}{0.36\times(15)} > \frac{1.2-1}{30\times1.2^2}$$

Which can be rearranged and simplified so as to read:

$$0.8\times1.2^2 > 0.072$$

As $1.2^2 > 1$, then it follows that 0.8 × $(>1) = (>0.8)$. And, hence, as $(>0.8) > 0.072$, given $\underline{m} = 0.6$ and $\overline{m} = 1.2$ the weakest form of conditional cooperation in the SDG implies a form of conditional cooperation in the CIG.

QED.

8.2.1.5. Spiteful preferences

8.2.1.5.1. Proof of proposition 4

Let's assume a subject's utility function, given g_i and g_j, is:

$$U_i^S(g_i, g_j) = \begin{cases} 30 - g_i + m \times (g_i + g_j) - \beta_i \times (g_j - g_i) \; if \; g_i \leq g_j \\ 30 - g_i + m \times (g_i + g_j) \; if \; g_i \geq g_j \end{cases}$$

Where $\beta_i \leq 0$. That is, a person with these preferences feels either pleasure or is indifferent at advantageous inequality ($\frac{\partial U_i(g_i, g_j)}{\partial (g_j - g_i)} = -\beta_i \geq 0$). These preferences represent someone who (i) derives pleasure from inequality provided that he is the one being better off in the distribution outcome. Otherwise, he does not feel any disadvantageous inequality. This is just the spiteful utility function $U_i^S(\pi_i, \pi_j)$ presented in chapter 4 once we substitute the material payoff function of the public goods game we are analysing.

Proposition 4. *If subject* i *maximizes the utility function* $U_i^S\left(\pi_i(g_i, g_j), \pi_j(g_i, g_j)\right)$, *where* i *contributes* g_i *and the other player contributes* g_j, *then subject* i*'s contribution attitudes, denoted as* c_i^*, *will be*

(i), in the Social Dilemma,

$$(\forall \beta_i), c_i^* = g_i = 0 \; \forall g_j \in A_j$$

(ii), in the Common Interest Game,

$$c_i^* = \begin{cases} g_i = 30 \; if \; g_j = 0 & \forall \beta_i \\ g_i = 0 \; \forall g_j \in \{10,20,30\} & iff \; \beta_i < \dfrac{30 \times (1 - \overline{m})}{g_j} \\ g_i = 30 \; \forall g_j \in \{10,20,30\} & iff \; \beta_i > \dfrac{30 \times (1 - \overline{m})}{g_j} \end{cases}$$

Proof.

The marginal derivative with respect to own contributions is:

$$\frac{\partial U_i^S\left(\pi_i(g_i,g_j),\pi_j(g_i,g_j)\right)}{\partial g_i}=\begin{cases}-1+m+\beta_i \; if \; g_i \leq g_j \\ \quad -1+m \; if \; g_i \geq g_j\end{cases}$$

For $\underline{m} < 1$, the second step of the marginal utility of own contributions is always negative. To see this, note $-1 + (< 1) = (< 0)$. The first step is negative when $\beta_i < 1 - \underline{m}$. For $\underline{m}$, it follows that $\beta_i < 1 - (< 1)$, as in the spiteful preferences model $\beta_i < 0$ and $1 - (< 1) = (> 0)$. Hence, the first derivative will be negative for all the values of $\underline{m} \in \left(\frac{1}{n}, 1\right)$. Given that the utility is linear in g_i and that the first derivative is negative alongside the whole domain of g_i for all values of $\underline{m}$, it follows that i's optimal cooperation attitudes in the SDG are given by:

$$(\forall \beta_i), c_i^* = g_i = 0 \; \forall \; g_j \in A_j$$

Which proves (i).

With regards to the CIG, the marginal derivative with respect to g_i is:

$$\frac{\partial U_i^S\left(\pi_i(g_i,g_j),\pi_j(g_i,g_j)\right)}{\partial g_i}=\begin{cases}-1+\overline{m}+\beta_i \; if \; g_i < g_j \\ -1+\overline{m} \qquad if \; g_i \geq g_j\end{cases}$$

For $\overline{m} \in (1, \infty)$, the second step of the marginal utility of own contributions is always positive. To see this, note that $\overline{m} > 1$. Hence, $-1 + (> 1) = (> 0)$. When $g_j = 0$, then $g_i \geq 0$. Hence, against $g_j = 0$ the best response is to fully contribute regardless of the value of β_i, as only the second step of the marginal derivative comes into play. This proves the first step of c_i^* in (ii).

Notice that the first step of the marginal derivative is negative when $\beta_i < 1 - \overline{m}$ and positive when $\beta_i > 1 - \overline{m}$.

This implies that, whenever $\beta_i > 1 - \overline{m}$, both steps of the marginal utility will be positive and, hence, full contribution against all contributions of the other player will be the best response, as the marginal derivative will be positive alongside the whole domain of g_i. Hence, it follows that

$$c_i^* = g_i = 30\ \forall g_j \in \{10{,}20{,}30\} \qquad\qquad iff\ \beta_i > \frac{30 \times (1-\overline{m})}{g_j}$$

Thereby proving the last step in c_i^* of (ii).

Additionally, notice that, whenever $\beta_i < 1-\overline{m}$, the first step of the marginal utility is negative. This implies that increasing contributions on the range $g_i < g_j$ decreases utility, thereby suggesting free riding as one potential optimal solution. The second step makes the marginal utility increasing in the range $g_i \geq g_j$, thereby suggesting full contribution as another potential optimal solution. Taken both results together, this indicates that we have two potential optimal best responses: free riding and full contribution.

Hence, person i's utility will be maximised by full contribution when $U_i^S(g_i = 30, g_j) > U_i^S(g_i = 0, g_j)$, which implies:

$$0 + \overline{m} \times (g_j + 30) > 30 + \overline{m} \times (g_j) - \beta_i \times (g_j)$$

Isolating β_i in the LHS and simplifying, we get:

$$\beta_i \times g_j > 30 + \overline{m} \times g_j - \overline{m} \times (g_j + 30)$$

Expanding the parenthesis of the RHS, we get:

$$\beta_i \times g_j > 30 + \overline{m} \times g_j - \overline{m} \times g_j - \overline{m} \times 30$$

Which, after simplifying, becomes:

$$\beta_i \times g_j > 30 - \overline{m} \times 30$$

And, taking 30 as a common factor in the RHS, we can rewrite the previous expression as:

$$\beta_i \times g_j > 30 \times (1-\overline{m})$$

And, hence,

$$\beta_i > \frac{30 \times (1 - \overline{m})}{g_j}$$

Whenever $g_j > 0$ and $\beta_i < 1 - \overline{m}$, $U_i^S(g_i = 30, g_j) > U_i^S(g_i = 0, g_j)$ will hold true whenever $\beta_i > \frac{30\times(1-MPCR)}{g_j}$, and $U_i^S(g_i = 30, g_j) < U_i^S(g_i = 0, g_j)$ whenever $\beta_i < \frac{30\times(1-MPCR)}{g_j}$. Therefore, the optimal contributions given the values of β_i are:

$$c_i^* = g_i = 30 \, \forall g_j \in \{10,20,30\} \qquad iff \; \beta_i > \frac{30 \times (1 - \overline{m})}{g_j}$$

$$c_i^* = g_i = 0 \, \forall g_j \in \{10,20,30\} \qquad iff \; \beta_i < \frac{30 \times (1 - \overline{m})}{g_j}$$

Which finishes proving (ii).

QED.

8.2.1.5.2. Other results involving spiteful preferences

Below we provide a corollary that presents the specific threshold values of β_i determining optimal contributions for each g_j.

Corollary 4.1. *If subject* i *maximizes the utility function* $U_i^S\left(\pi_i(g_i, g_j), \pi_j(g_i, g_j)\right)$, *and given* $\underline{m} = 0.6$ *and* $\overline{m} = 1.2$, *the subject* i*'s choices will*

(i), in the Social Dilemma, be

$$(\forall\beta_i), c_i^* = g_i = 0 \;\forall g_j \in A_j$$

(ii), in the Common Interest Game, be

(a) $(\forall\beta_i), g_i = 30 \; if \; g_j = 0$
(b) $g_i = 0$ *against* $g_j = 10$ *if* $\beta_i < -0.6$
(c) $g_i = 30$ *against* $g_j = 10$ *if* $\beta_i > -0.6$
(d) $g_i = 0$ *against* $g_j = 20$ *if* $\beta_i < -0.3$
(e) $g_i = 30$ *against* $g_j = 20$ *if* $\beta_i > -0.3$
(f) $g_i = 0$ *against* $g_j = 30$ *if* $\beta_i < -0.2$
(g) $g_i = 30$ *against* $g_j = 30$ *if* $\beta_i > -0.2$

Proof.

Part (i) trivially follows from proposition 4, and therefore needs no proof.

Regarding part (ii), recall the last two conditions found in proposition 4:

$$c_i^* = g_i = 30 \;\forall g_j \in \{10,20,30\} \qquad iff \; \beta_i > \frac{30 \times (1 - \overline{m})}{g_j}$$

$$c_i^* = g_i = 0 \;\forall g_j \in \{10,20,30\} \qquad iff \; \beta_i < \frac{30 \times (1 - \overline{m})}{g_j}$$

Substituting $\overline{m} = 1.2$ and simplifying, we get:

$$c_i^* = g_i = 30 \;\forall g_j \in \{10,20,30\} \qquad iff \; \beta_i > \frac{-6}{g_j}$$

$$c_i^* = g_i = 0 \; \forall g_j \in \{10,20,30\} \qquad iff \; \beta_i < \frac{-6}{g_j}$$

Substituting for all values of $g_i \in \{10,20,30\}$, we get the following conditions:

$c_i^* = g_i = 0$ against $g_j = 10$ *iff* $\beta_i < -0.6$
$c_i^* = g_i = 30$ against $g_j = 10$ *iff* $\beta_i > -0.6$
$c_i^* = g_i = 0$ against $g_j = 20$ *iff* $\beta_i < -0.3$
$c_i^* = g_i = 30$ against $g_j = 20$ *iff* $\beta_i > -0.3$
$c_i^* = g_i = 0$ against $g_j = 30$ *iff* $\beta_i < -0.2$
$c_i^* = g_i = 30$ against $g_j = 30$ *iff* $\beta_i > -0.2$
QED.

8.2.1.6. Social Efficiency preferences

8.2.1.6.1. Proof of proposition 5

Proposition 5. *If subject i maximizes the utility function* $U_i^{SE}\left(\pi_i(g_i, g_j), \pi_j(g_i, g_j)\right)$*, where* i *contributes* g_i *and the other player contributes* g_j*, then subject* i*'s contribution attitudes, denoted as* c_i^**, will be*

(i), in the Social Dilemma,

(a) $c_i^* = g_i = 0\ \forall g_j \in A_j\ iff\ p_i < \frac{1-m}{m}$

(b) $c_i^* = g_i \in A_i\ \forall g_j \in A_j\ iff\ p_i = \frac{1-m}{m}$

(c) $c_i^* = g_i = 30\ \forall g_j \in A_j\ iff\ p_i > \frac{1-m}{m}$

(ii), in the Common Interest Game,

$$(\forall \beta_i), c_i^* = g_i = 30\ \forall g_j \in A_j$$

Proof.

Let's start by writing the utility function of person i for generic levels of contribution g_i and g_j:

$$U_i^{SE}\left(\pi_i(g_i, g_j), \pi_j(g_i, g_j)\right) = (1 - p_i) \times \pi_i(g_i, g_j) + p_i \times \left(\pi_i(g_i, g_j) + \pi_j(g_i, g_j)\right)$$

Expanding the RHS, we get:

$$U_i^{SE}\left(\pi_i(g_i, g_j), \pi_j(g_i, g_j)\right)$$
$$= \pi_i(g_i, g_j) - p_i \times \pi_i(g_i, g_j) + p_i \times \pi_i(g_i, g_j) + p_i \times \pi_j(g_i, g_j)$$

Given that $-p_i \times \pi_i(g_i, g_j) + p_i \times \pi_i(g_i, g_j) = 0$ and simplifying, we get:

$$U_i^{SE}\left(\pi_i(g_i,g_j),\pi_j(g_i,g_j)\right) = \pi_i(g_i,g_j) + p_i \times \pi_j(g_i,g_j)$$

Substituting both $\pi_i(g_i,g_j)$ and $\pi_j(g_i,g_j)$ by the material payoff function defined in chapter 4, we get:

$$U_i^{SE}\left(\pi_i(g_i,g_j),\pi_j(g_i,g_j)\right) = 30 - g_i + m \times (g_i + g_j) + p_i \times \{30 - g_j + m \times (g_i + g_j)\}$$

Once we have expressed the utility of person i explicitly in terms of g_i and g_j, we can calculate the marginal utility with respect to g_i to see whether person i increases or decreases his or her utility in his or her own contributions:

$$\frac{\partial U_i^{SE}\left(\pi_i(g_i,g_j),\pi_j(g_i,g_j)\right)}{\partial g_i} = -1 + m + p_i \times m$$

Note that, whenever $\overline{m} \in (1,\infty)$, the marginal utility becomes:

$$\frac{\partial U_i^{SE}\left(\pi_i(g_i,g_j),\pi_j(g_i,g_j)\right)}{\partial g_i} = -1 + (>1) \times (1 + p_i)$$

Given that $p_i \in [0,1]$, the marginal utility will always be positive, as:

$$\frac{\partial U_i^{SE}}{\partial g_i} = -1 + (>1) \times \left(1 + (\geq 0)\right) = -1 + (>1) \times (\geq 1) = -1 + (>1) = (>0)$$

Hence, the best response for a common interest game, where $\overline{m} \in (1,\infty)$ is given by:

$$(\forall\, p_i[0,1])\,, c_i^* = \; g_i = 30\; \forall g_j \in A_j$$

Which proves (ii).

In a social dilemma, where $\underline{m} \in \left(\frac{1}{n}, 1\right)$, the value of the marginal utility can be positive or negative depending on the value of p_i. To find for which values of p_i does the marginal utility of $g_i$becomes negative, we just isolate p_i in the LHS of the marginal utility found above to get:

$$p_i \times \underline{m} < 1 - \underline{m}$$

Which, dividing both hand sides by $\underline{m}$, becomes:

$$p_i < \frac{1 - \underline{m}}{\underline{m}}$$

Hence, when $p_i < \frac{1-\underline{m}}{\underline{m}}$ (resp. $p_i > \frac{1-\underline{m}}{\underline{m}}$) the utility of person i decreases (resp. increases) as he or she increases (resp. decreases) his or her contributions. Hence, the best response is given by:

$$c_i^* = \begin{cases} g_i = 0 \; \forall g_j \in A_j & if \; p_i < \dfrac{1 - \underline{m}}{\underline{m}} \\ g_i \in A_i \; \forall g_j \in A_j & if \; p_i = \dfrac{1 - \underline{m}}{\underline{m}} \\ g_i = 30 \; \forall g_j \in A_j & if \; p_i > \dfrac{1 - \underline{m}}{\underline{m}} \end{cases}$$

Which proves all points in (i).

QED.

8.2.1.6.2. Other results involving social efficiency preferences

Below we provide a corollary that presents the specific threshold values of p_i determining optimal contributions for each g_j.

Corollary 5.1.: *If subject* i *maximizes the utility function* $U_i^{SE}\left(\pi_i(g_i, g_j), \pi_j(g_i, g_j)\right)$*, and given* $\underline{m} = 0.6$ *and* $\overline{m} = 1.2$*, the subject* i*'s choices will*

(i), in the Social Dilemma, be

$$c_i^* = \begin{cases} g_i = 0\,\forall g_j \in A_j & if\ p_i < \frac{2}{3} \\ g_i \in A_i\,\forall g_j \in A_j & if\ p_i = \frac{2}{3} \\ g_i = 30\,\forall g_j \in A_j & if\ p_i > \frac{2}{3} \end{cases}$$

(ii), in the Common Interest Game, be

$$(\forall p_i), g_i = 30\ \forall\ g_j \in A_j$$

Proof.

(a) Given the best response for the social dilemma found in proposition 5, and substituting $\underline{m} = 0.6$, we get:

$$c_i^* = \begin{cases} g_i = 0\,\forall g_j \in A_j & if\ p_i < \frac{2}{3} \\ g_i \in A_i\,\forall g_j \in A_j & if\ p_i = \frac{2}{3} \\ g_i = 30\,\forall g_j \in A_j & if\ p_i > \frac{2}{3} \end{cases}$$

Which proves (i). Point (ii) is self-evident given proposition 5.

QED.

8.2.1.7. Maximin preferences

8.2.1.7.1. Proof of proposition 6

Proposition 6. *If subject i maximizes the utility function* $U_i^{MM}\left(\pi_i(g_i, g_j), \pi_j(g_i, g_j)\right)$*, where* i *contributes* g_i *and the other player contributes* g_j*, then subject* i*'s contribution attitudes, denoted as* c_i^**, will be*

(i), in the Social Dilemma,

(a) $c_i^* = g_i = 0\ \forall g_j \in A_j\ iff\ q_i < 1 - \underline{m}$
(b) $c_i^* = g_i \in [0, g_j]\ \forall g_j \in A_j\ iff\ q_i = 1 - \underline{m}$
(c) $c_i^* = g_i = g_j\ \forall g_j \in A_j\ iff\ q_i > 1 - \underline{m}$

(ii), in the Common Interest Game,

$$(\forall \beta_i), c_i^* = g_i = 30\ \forall g_j \in A_j$$

Proof.

Let's start by writing the utility function of person i for generic levels of contribution g_i and g_j:

$$U_i^{MM}\left(\pi_i(g_i, g_j), \pi_j(g_i, g_j)\right) = (1 - q_i) \times \pi_i(g_i, g_j) + q_i \times min\{\pi_i(g_i, g_j), \pi_j(g_i, g_j)\}$$

Using the results of Lemma 0 (a), we know that $min\{\pi_i(g_i, g_j), \pi_j(g_i, g_j)\} = \pi_i(g_i, g_j)$ whenever $g_i > g_j$ and $min\{\pi_i(g_i, g_j), \pi_j(g_i, g_j)\} = \pi_j(g_i, g_j)$ whenever $g_i < g_j$. Hence, we can rewrite the previous utility function as follows:

$$U_i\left(\pi_i(g_i, g_j), \pi_j(g_i, g_j)\right) = \begin{cases} (1 - q_i) \times \pi_i(g_i, g_j) + q_i \times \pi_i(g_i, g_j)\ if\ g_i \geq g_j \\ (1 - q_i) \times \pi_i(g_i, g_j) + q_i \times \pi_j(g_i, g_j)\ if\ g_i < g_j \end{cases}$$

By taking $\pi_i(g_i, g_j)$ as a common factor when $g_i \geq g_j$ and expanding the first parenthesis when $g_i < g_j$, we get:

$$U_i\left(\pi_i(g_i, g_j), \pi_j(g_i, g_j)\right) = \begin{cases} \pi_i(g_i, g_j) \times (1 - q_i + q_i) \; if \; g_i \geq g_j \\ \pi_i(g_i, g_j) - q_i \times \pi_i(g_i, g_j) + q_i \times \pi_j(g_i, g_j) \; if \; g_i < g_j \end{cases}$$

Simplifying when $g_i \geq g_j$ and taking q_i as a common factor when $g_i < g_j$, we get:

$$U_i\left(\pi_i(g_i, g_j), \pi_j(g_i, g_j)\right) = \begin{cases} \pi_i(g_i, g_j) \; if \; g_i \geq g_j \\ \pi_i(g_i, g_j) + q_i \times \left(\pi_j(g_i, g_j) - \pi_i(g_i, g_j)\right) \; if \; g_i < g_j \end{cases}$$

Using Lemma 0 (b), we can substitute $\pi_j(g_i, g_j) - \pi_i(g_i, g_j) = g_i - g_j$ when $g_i < g_j$ to get:

$$U_i\left(\pi_i(g_i, g_j), \pi_j(g_i, g_j)\right) = \begin{cases} \pi_i(g_i, g_j) & if \; g_i \geq g_j \\ \pi_i(g_i, g_j) + q_i \times (g_i - g_j) & if \; g_i < g_j \end{cases}$$

Substituting $\pi_i(g_i, g_j)$ by the corresponding material payoff function outlined above, we get:

$$U_i\left(\pi_i(g_i, g_j), \pi_j(g_i, g_j)\right) = \begin{cases} 30 - g_i + m \times (g_i + g_j) & if \; g_i \geq g_j \\ 30 - g_i + m \times (g_i + g_j) + q_i \times (g_i - g_j) & if \; g_i < g_j \end{cases}$$

Taking the marginal derivative of person i's utility function with respect to his or her own contributions, we get:

$$\frac{\partial U_i\left(\pi_i(g_i, g_j), \pi_j(g_i, g_j)\right)}{\partial g_i} = \begin{cases} -1 + m \; if \; g_i \geq g_j \\ -1 + m + q_i \; if \; g_i < g_j \end{cases}$$

(a)

Note that, in a common interest game, where $\overline{m} \in (1, \infty)$, the marginal derivative of person i's utility function becomes positive regardless of the value of g_i. To see this, note that the first step takes the following values:

$$-1 + (> 1) = (> 0)$$

Given that $q_i \in [0,1]$, the second step takes the following values:

$$-1 + (> 1) + (\geq 0) = (> 0)$$

Hence, the optimal contribution for person i in the CIG becomes:

$$c_i^* = g_j = 30\ \forall g_j \in A_j\ \forall q_i \in [0,1]$$

Which proves (ii).

In a social dilemma game, where $\underline{m} \in \left(\frac{1}{n}, 1\right)$, the marginal derivative of person i's utility function becomes negative regardless of the value of q_i when $g_i \geq g_j$. This is so as $-1 + \underline{m}$ if always negative for $\underline{m} < 1$.

The marginal utility of own contribution when $g_i < g_j$ depends on the value of q_i. More specifically, the marginal utility will be positive in that range whenever the following inequality holds true:

$$-1 + \underline{m} + q_i > 0$$

Which implies the condition $q_i > 1 - \underline{m}$. Hence, when $q_i > 1 - \underline{m}$ a person will find it profitable to increase his contributions whenever $g_i < g_j$, and unprofitable to keep increasing his contributions in the range $g_i \geq g_j$. It, then, follows that the best response when $q_i > 1 - \underline{m}$ is to contribute $g_i = g_j$:

$$c_i^* = g_i = g_j\ \forall g_j \in A_j\ iff\ q_i > 1 - \underline{m}$$

Following an analogous logic, the best response when $q_i < 1 - \underline{m}$ is to contribute $g_i = 0$ for all g_j; as, subject to those parameter values, increasing contributions decreases utility in the range $g_i < g_j$. Hence,

$$c_i^* = g_i = 0\ \forall g_j \in A_j\ iff\ q_i < 1 - \underline{m}$$

Finally, whenever $q_i = 1 - \underline{m}$, a person will be indifferent between any g_i in the range $[0, g_j]$, as the marginal utility does not vary with own contributions in this case.

More compactly, one can express those results as follows:

$$c_i^* = \begin{cases} g_i = 0\ \forall g_j \in A_j & if\ q_i < 1 - \underline{m} \\ g_i \in [0, g_j]\ \forall g_j \in A_j & if\ q_i = 1 - \underline{m} \\ g_i = g_j\ \forall g_j \in A_j & if\ q_i > 1 - \underline{m} \end{cases}$$

Which proves (i).

QED.

8.2.1.7.2. Other results involving maximin preferences

Below we provide a corollary that presents the specific threshold values of q_i determining optimal contributions for each g_j.

Corollary 6.1. *If subject* i *maximizes the utility function* $U_i^{MM}\left(\pi_i(g_i, g_j), \pi_j(g_i, g_j)\right)$, *and given* $\underline{m} = 0.6$ *and* $\overline{m} = 1.2$, *the subject* i*'s choices will*

(i), in the Social Dilemma, be

$$c_i^* = \begin{cases} g_i = 0\ \forall g_j \in A_j & if\ q_i < 0.4 \\ g_i \in A_i\ \forall g_j \in A_j & if\ q_i = 0.4 \\ g_i = 30\ \forall g_j \in A_j & if\ q_i > 0.4 \end{cases}$$

(ii), in the Common Interest Game, be

$$(\ \forall q_i), g_i = 30\ \forall\ g_j \in \{0,10,20,30\}$$

Proof.

Given the best response for the social dilemma found in proposition 6, and substituting $\underline{m} = 0.6$, we get:

$$c_i^* = \begin{cases} g_i = 0\ \forall g_j \in A_j & if\ q_i < 0.4 \\ g_i \in A_i\ \forall g_j \in A_j & if\ q_i = 0.4 \\ g_i = 30\ \forall g_j \in A_j & if\ q_i > 0.4 \end{cases}$$

Which proves (i). Point (ii) is self-evident given proposition 6.

QED.

8.2.2. Proofs regarding estimated parameters through the use of parameter-elicitation games

8.2.2.1. Ultimatum Games

In the derivations below, we use the following notation:

- $x \in [0,7]$ represents the offer made by the sender

- 14 is the initial endowment of the sender

- 0 is the quantity that both get if the receiver rejects the sender's offer

- ε is an arbitrarily small number representing the smallest increase and or decrease of an offer.

- i is referred to as the receiver, and hence $U_i()$ represents the utility of the receiver

- A given distribution $(x, 14 - x)$ represents the payoff of the receiver in the first place (x) and the payoff of the sender in the second place $(14 - x)$. That is, we define $\pi_i(x, 14 - x) = x$ and $\pi_j(x, 14 - x) = 14 - x$.

8.2.2.1.1. Disadvantageous Inequality parameter

8.2.2.1.1.1. Proof of proposition 7

Proposition 7. *If subject i maximizes the utility function* $U_i^{FS}\left(\pi_i(x, 14-x), \pi_j(x, 14-x)\right)$, *subject i's minimum acceptable offer is* $x+\varepsilon+\varepsilon$ *and subject i's maximum rejectable offer is* $x+\varepsilon$, *where* $x+\varepsilon+\varepsilon \leq 7$ *and* $x+\varepsilon \geq 0$, *then subject i's choices would reveal an* α_i *parameter between the following boundaries:*

$$\frac{x+\varepsilon}{14-2\times(x+\varepsilon)} < \alpha_i < \frac{x+\varepsilon+\varepsilon}{14-2\times(x+\varepsilon+\varepsilon)}$$

Proof.

As a generic offer $x \in [0,7]$, then it follows that $14-x \in [7,14]$. Hence, $14-x \geq x$, and no offer goes above 7 regardless of the value of ε. This means that $U_i^{FS}\left(\pi_i(x, 14-x), \pi_j(x, 14-x)\right)$ will be on the domain of disadvantageous inequality as $14-x \geq x$ implies $\pi_i(x, 14-x) < \pi_j(x, 14-x)$. Hence, $U_i^{FS}\left(\pi_i(x, 14-x), \pi_j(x, 14-x)\right)$for the 2-person ultimatum game described above, for a generic offer x, is:

$$U_i^{FS}(x, 14-x) = 14-x-\alpha_i \times (14-x-x)$$

To compute the generic threshold of α_i, we assume a person's minimum acceptable offer is $x+\varepsilon+\varepsilon$ and his or her maximum rejectable offer is $x+\varepsilon$ as stated in the proposition, where $\varepsilon \geq 0$, and $x+\varepsilon+\varepsilon \leq 7$. This would imply that the utility of accepting the minimum acceptable offer is greater than the utility of the distribution $(0,0)$ and that the utility of accepting the maximum rejectable offer is lower than the utility of the distribution $(0,0)$. In mathematical terms:

$$U_i^{FS}(x+\varepsilon, 14-x-\varepsilon) < U_i^{FS}(0,0)$$
$$U_i^{FS}(x+\varepsilon+\varepsilon, 14-x-\varepsilon-\varepsilon) > U_i^{FS}(0,0)$$

Substituting the generic utility function by the Fehr-Schmidt specification presented in chapter 4, the two equations above would transform into:

$$x + \varepsilon - \alpha_i \times \big(14 - x - \varepsilon - (x + \varepsilon)\big) < 0$$

$$x + \varepsilon + \varepsilon - \alpha_i \times \big(14 - x - \varepsilon - \varepsilon - (x + \varepsilon + \varepsilon)\big) > 0$$

Simplifying, we get:

$$x + \varepsilon - \alpha_i \times \big(14 - 2 \times (x + \varepsilon)\big) < 0$$

$$x + \varepsilon + \varepsilon - \alpha_i \times \big(14 - 2 \times (x + \varepsilon + \varepsilon)\big) > 0$$

Which collapse to:

$$\alpha_i > \frac{(x + \varepsilon)}{14 - 2 \times (x + \varepsilon)}$$

$$\alpha_i < \frac{(x + \varepsilon + \varepsilon)}{14 - 2 \times (x + \varepsilon + \varepsilon)}$$

And, hence, it follows that:

$$\frac{(x + \varepsilon)}{14 - 2 \times (x + \varepsilon)} < \alpha_i < \frac{(x + \varepsilon + \varepsilon)}{14 - 2 \times (x + \varepsilon + \varepsilon)}$$

QED.

8.2.2.1.1.2. More proofs on the disadvantageous inequality parameter elicitation

As we showed in corollary 2.1 (b), the key value of the disadvantageous inequality parameter for our predictions of inequality aversion preferences regarding cooperation attitudes in the CIG is $\alpha_i \gtreqless 0.2$. Below we provide a corollary showing that a minimum acceptable offer (resp. maximum rejectable offer) of $x = 2$ precisely reveals this threshold.

Corollary 7.1. *Let's suppose that subject* i *maximizes the utility function* $U_i^{FS}\left(\pi_i(x, 14 - x), \pi_j(x, 14 - x)\right)$. *Then, if subject* i*'s minimum acceptable offer is 2 or lower subject* i *reveals* $\alpha_i < 0.2$. *If subject* i*'s maximum rejectable offer is 2 or higher subject* i *reveals* $\alpha_i > 0.2$

Given the inequalities found in Proposition 7, it follows that a minimum acceptable offer of 2 or lower would entail:

$$\alpha_i < \frac{(\leq 2)}{\left(14 - 2 \times \left((\leq 2)\right)\right)}$$

Similarly, a maximum rejectable offer of 2 or higher would entail:

$$\alpha_i > \frac{(\geq 2)}{\left(14 - 2 \times \left((\geq 2)\right)\right)}$$

And, hence,

$$\alpha_i < \frac{(\leq 2)}{\left(14 - (\leq 4)\right)}$$

$$\alpha_i > \frac{(\geq 2)}{\left(14 - (\geq 4)\right)}$$

Which becomes:

$$\alpha_i < \frac{(\leq 2)}{(\geq 10)}$$

$$\alpha_i > \frac{(\geq 2)}{(\leq 10)}$$

Now, let’s define a partially ordered set:

$$P := (X, \leq)$$

Where

$$X \coloneqq \{x \in X | x \geq 0\}$$

We define the set MAO as follows:

$$MAO \coloneqq \left\{x \in MAO| \left((x \in X) \wedge \left(x < \frac{(\leq 2)}{(\geq 10)}\right)\right)\right\}$$

And, also, we define the set MRO as follows:

$$MRO \coloneqq \left\{x \in MRO| \left((x \in X) \wedge \left(x > \frac{(\geq 2)}{(\leq 10)}\right)\right)\right\}$$

Where MAO stands for ‘*Minimum Acceptable Offer*’ and MRO stands for ‘*Maximum Rejectable Offer*’. It is straightforward to see that MAO is bounded above by $y = \frac{2}{10}$, as (i) $y \geq x \; \forall x \in MAO$ and (ii) $y \geq 0$ and, hence, $y \in X$.

Using a similar logic, it is also straightforward to see that MRO is bounded below by $y = \frac{2}{10}$, as (i) $x \geq y \; \forall x \in MRO$ and (ii) y$\geq$ 0 and, hence, $y \in X$.

Given that $y = \frac{2}{10}$ is an upper (lower) bound of MAO (MRO), and that it is the lowest upper bound (greatest lower bound) of MAO (MRO), it trivially follows that:

$$maxMAO = supMAO = \frac{2}{10} \in X$$

$$minMRO = infMRO = \frac{2}{10} \in X$$

It, then, follows, that the values of α_i for the first (second) inequality found above must be lower than the supremum of MAO (greater than the infimum of MRO):

$$\alpha_i < SupMAO$$

$$\alpha_i > InfMAO$$

And, substituting the values of $supMAO$ and $infMRO$, we get:

$$\alpha_i < 0.2$$

$$\alpha_i > 0.2$$

It follows that a person whose minimum acceptable offer is 2 or lower reveals $\alpha_i < 0.2$ and a person whose maximum rejectable offer is 2 or higher reveals $\alpha_i > 0.2$

QED.

8.2.2.2. Modified Dictator Games

In the derivations below, we use the following notation:

- $(20{,}0)$ is the original allocation that the dictator can choose instead of the equitable allocation

- $x \in [0{,}32]$ refers to the value that each gets from the equitable allocation. Hence, a given distribution (x, x) represents the payoff of the dictator and the receiver.

- ε is an arbitrarily small number representing the smallest increase and or decrease in the value each gets from the equitable allocation.

- i is referred to as the dictator, and hence $U_i()$ represents the utility of the dictator

8.2.2.2.1. Advantageous Inequality and Spiteful parameters

8.2.2.2.1.1. Proof of proposition 8

Proposition 8. *If subject* i *that maximizes the utility function* $U_i^{FS}(\pi_i, \pi_j)$*, whose maximum rejection quantity is* $x + \varepsilon$*, from the distribution* $(x + \varepsilon, x + \varepsilon)$*, to accept a distribution* $(20{,}0)$*, and whose minimum accepting quantity is* $x + \varepsilon + \varepsilon$*, from the distribution* $(x + \varepsilon + \varepsilon, x + \varepsilon + \varepsilon)$*, to reject a distribution* $(20{,}0)$*, will have a* β_i *parameter within the following boundaries:*

$$\frac{20 - (x + \varepsilon + \varepsilon)}{20} < \beta_i < \frac{20 - (x + \varepsilon)}{20}$$

Proof.

Let's assume a person with $U_i^{FS}(\pi_i, \pi_j)$ reveals the following preference pattern with their choices in the modified dictator games:

$$U_i^{FS}(20{,}0) > U_i^{FS}(x + \varepsilon, x + \varepsilon)$$

$$U_i^{FS}(20{,}0) < U_i^{FS}(x + \varepsilon + \varepsilon, x + \varepsilon + \varepsilon)$$

Substituting the generic utility by the inequality aversion preferences, the equations can be rewritten as:

$$20 - \beta_i \times (20) > x + \varepsilon$$

$$20 - \beta_i \times (20) < x + \varepsilon + \varepsilon$$

Isolating β_i in the RHS, we get:

$$20 - (x + \varepsilon) > \beta_i \times (20)$$

$$20 - (x + \varepsilon + \varepsilon) < \beta_i \times (20)$$

Which simplify to:

$$\frac{20 - (x + \varepsilon)}{20} > \beta_i$$

$$\frac{20 - (x + \varepsilon + \varepsilon)}{20} < \beta_i$$

Hence, β_i can be expressed in terms of the two thresholds together:

$$\frac{20 - (x + \varepsilon + \varepsilon)}{20} < \beta_i < \frac{20 - (x + \varepsilon)}{20}$$

QED.

8.2.2.2.1.2. More proofs on the advantageous inequality and spiteful parameters elicitation

As we showed in corollary 2.1 (a), the key value of the advantageous inequality parameter for our predictions of inequality aversion preferences regarding cooperation attitudes in the SDG is $\beta_i \gtreqless 0.4$. Also, corollary 4.1 showed that the relevant parameter values of β_i for play in the CIG were $\beta_i \gtreqless -0.6$, $\beta_i \gtreqless -0.3$, and $\beta_i \gtreqless -0.2$. Below we provide a corollary showing that a maximum rejecting quantity (resp. minimum accepting quantity) of $x = 12$ reveals the necessary threshold for the inequality aversion model, and that a maximum rejecting quantity (resp. minimum accepting quantity) of $x = 24$, $x = 26$ and $x = 32$ reveal the necessary thresholds for predictions of cooperation attitudes in the CIG for the spiteful preferences model.

Corollary 8.1. *Let's suppose that subject* i *maximizes the utility function* $U_i^{FS}(\pi_i, \pi_j)$*. Then,*

(a) If subject i*'s minimum accepting quantity is 12 or lower subject* i *reveals* $\beta_i > 0.4$*. If subject* i*'s maximum rejecting quantity is 12 or higher subject* i *reveals* $\beta_i < 0.4$*.*

(b) If subject i*'s minimum accepting quantity is 24 or lower subject* i *reveals* $\beta_i > -0.2$*. If subject* i*'s maximum rejecting quantity is 24 or higher subject* i *reveals* $\beta_i < -0.2$*.*

(c) If subject i*'s minimum accepting quantity is 26 or lower subject* i *reveals* $\beta_i > -0.3$*. If subject* i*'s maximum rejecting quantity is 26 or higher subject* i *reveals* $\beta_i < -0.3$*.*

(d) If subject i*'s minimum accepting quantity is 32 or lower subject* i *reveals* $\beta_i > -0.6$*. . If subject* i*'s maximum rejecting quantity is 32 or higher subject* i *reveals* $\beta_i < -0.6$*.*

Proof.

(a)

Given the inequality found in Proposition 8, it follows that a minimum accepting quantity of 12 or lower would entail:

$$\beta_i > \frac{20 - (\leq 12)}{20}$$

Similarly, a maximum rejecting quantity of 2 or higher would entail:

$$\beta_i < \frac{20 - (\geq 12)}{20}$$

And, hence,

$$\beta_i > \frac{\geq 8}{20}$$

$$\beta_i < \frac{\leq 8}{20}$$

Now, let's define a partially ordered set:

$$P: = (X, \leq)$$

Where

$$X \coloneqq \{x \in X | x \geq 0\}$$

We define the set MAQ as follows:

$$MAQ \coloneqq \left\{x \in MAQ | \left((x \in X) \wedge \left(x > \frac{\geq 8}{20}\right)\right)\right\}$$

And, also, we define the set MRO as follows:

$$MRQ \coloneqq \left\{x \in MRQ | \left((x \in X) \wedge \left(x < \frac{\leq 8}{20}\right)\right)\right\}$$

Where MAQ stands for '*Minimum Accepting Quantity*' and MRO stands for '*Maximum Rejecting Quantity*'. It is straightforward to see that MAQ is bounded below by $y = \frac{8}{20}$, as (i) $y \leq x \; \forall x \in MAQ$ and (ii) $y \geq 0$ and, hence, $y \in X$.

Using a similar logic, it is also straightforward to see that MRQ is bounded above by $y = \frac{8}{20}$, as (i) $y \geq x \; \forall x \in MRQ$ and (ii) y$\geq$ 0 and, hence, $y \in X$.

Given that $y = \frac{8}{20}$ is a lower (upper) bound of MAQ (MRQ), and that it is the greatest lower bound (least upper bound) of MAQ (MRQ), it trivially follows that:

$$minMAQ = infMAQ = \frac{8}{20} \in X$$

$$maxMRQ = supMRQ = \frac{8}{20} \in X$$

It, then, follows, that the values of β_i for the first (second) inequality found above must be greater than the infimum of MAQ (lower than the supremum of MRQ):

$$\beta_i > infMAQ$$

$$\beta_i < supMRQ$$

And, substituting the values of $infMAQ$ and $supMRQ$, we get:

$$\beta_i > 0.4$$

$$\beta_i < 0.4$$

It follows that a person whose minimum accepting quantity is 12 or lower reveals $\beta_i > 0.4$ and a person whose maximum rejecting quantity is 12 or higher reveals $\beta_i < 0.4$

(b)

Following (a), a minimum accepting quantity of 24 or lower and a maximum rejecting quantity of 24 or higher would entail:

$$\beta_i > \frac{20 - (\leq 24)}{20}$$

$$\beta_i < \frac{20 - (\geq 24)}{20}$$

And, hence,

$$\beta_i > \frac{-(\leq 4)}{20}$$

$$\beta_i < \frac{-(\geq 4)}{20}$$

Now, let's define a partially ordered set:

$$P{:} = (X, \leq)$$

Where

$$X \coloneqq \{x \in X | x \leq 0\}$$

We define the set MAQ as follows:

$$MAQ \coloneqq \left\{x \in MAQ | \left((x \in X) \wedge \left(x > \frac{-(\leq 4)}{20}\right)\right)\right\}$$

And, also, we define the set MRO as follows:

$$MRQ \coloneqq \left\{x \in MRQ | \left((x \in X) \wedge \left(x < \frac{-(\geq 4)}{20}\right)\right)\right\}$$

It is straightforward to see that MAQ is bounded below by $y = -\frac{4}{20}$, as (i) $y \leq x\ \forall x \in MAQ$ and (ii) $y \leq 0$ and, hence, $y \in X$.

Using a similar logic, it is also straightforward to see that MRQ is bounded above by $y = \frac{4}{20}$, as (i) $y \geq x\ \forall x \in MRQ$ and (ii) y≤ 0 and, hence, $y \in X$.

Given that $y = -\frac{4}{20}$ is a lower (upper) bound of MAQ (MRQ), and that it is the greatest lower bound (least upper bound) of MAQ (MRQ), it trivially follows that:

$$minMAQ = infMAQ = -\frac{4}{20} \in X$$

$$maxMRQ = supMRQ = -\frac{4}{20} \in X$$

It, then, follows, that the values of β_i for the first (second) inequality found above must be greater than the infimum of MAQ (lower than the supremum of MRQ):

$$\beta_i > infMAQ$$

$$\beta_i < supMRQ$$

And, substituting the values of $infMAQ$ and $supMRQ$, we get:

$$\beta_i > -0.2$$

$$\beta_i < -0.2$$

It follows that a person whose minimum accepting quantity is 24 or lower reveals $\beta_i > -0.2$ and a person whose maximum rejecting quantity is 24 or higher reveals $\beta_i < -0.2$

(c)

Following (b), a minimum accepting quantity of 26 or lower and a maximum rejecting quantity of 26 or higher would entail:

$$\beta_i > \frac{20 - (\leq 26)}{20}$$

$$\beta_i < \frac{20 - (\geq 26)}{20}$$

And, hence,

$$\beta_i > \frac{-(\leq 6)}{20}$$

$$\beta_i < \frac{-(\geq 6)}{20}$$

Using the same technique as in (b), which we omit to avoid unnecessary repetition, it follows that:

$$\beta_i > -0.3$$

$$\beta_i < -0.3$$

It follows that a person whose minimum accepting quantity is 26 or lower reveals $\beta_i > -0.3$ and a person whose maximum rejecting quantity is 26 or higher reveals $\beta_i < -0.3$

(d)

Following (b), a minimum accepting quantity of 32 or lower and a maximum rejecting quantity of 32 or higher would entail:

$$\beta_i > \frac{20 - (\leq 32)}{20}$$

$$\beta_i < \frac{20 - (\geq 32)}{20}$$

And, hence,

$$\beta_i > \frac{-(\leq 12)}{20}$$

$$\beta_i < \frac{-(\geq 12)}{20}$$

Using the same technique as in (b), which we omit to avoid unnecessary repetition, it follows that:

$$\beta_i > -0.6$$

$$\beta_i < -0.6$$

It follows that a person whose minimum accepting quantity is 32 or lower reveals $\beta_i > -0.6$ and a person whose maximum rejecting quantity is 32 or higher reveals $\beta_i < -0.6$

QED.

8.2.2.2.2. Social Efficiency parameter

8.2.2.2.2.1. Proof of proposition 9.

Proposition 9. *Let's suppose that subject* i *maximizes the utility function* $U_i^{SE}(\pi_i, \pi_j)$. *If subject* i*'s maximum rejection quantity is* $x+\varepsilon$, *from the distribution* $(x+\varepsilon, x+\varepsilon)$, *to accept a distribution* $(20{,}0)$, *and if subject* i*'s minimum accepting quantity is* $x+\varepsilon+\varepsilon$, *from the distribution* $(x+\varepsilon+\varepsilon, x+\varepsilon+\varepsilon)$, *to reject a distribution* $(20{,}0)$, *then subject* i *will reveal to have a* p_i *parameter within the following boundaries:*

$$\frac{20-(x+\varepsilon+\varepsilon)}{x+\varepsilon+\varepsilon} < p_i < \frac{20-(x+\varepsilon)}{x+\varepsilon}$$

Proof.

Let's assume a person with $U_i^{SE}\left(\pi_i(x, 14-x), \pi_j(x, 14-x)\right)$ preferences reveals the following preference pattern with their choices in the modified dictator games:

$$U_i^{SE}(20{,}0) > U_i^{SE}(x+\varepsilon, x+\varepsilon)$$

$$U_i^{SE}(20{,}0) < U_i^{SE}(x+\varepsilon+\varepsilon, x+\varepsilon+\varepsilon)$$

These equations can be rewritten as:

$$(1-p_i)\times 20 + p_i \times (20) > (1-p_i)\times(x+\varepsilon) + p_i \times (x+\varepsilon+x+\varepsilon)$$

$$(1-p_i)\times 20 + p_i \times (20) < (1-p_i)\times(x+\varepsilon+\varepsilon) + p_i \times (x+\varepsilon+\varepsilon+x+\varepsilon+\varepsilon)$$

Which, by taking 20 as a common factor in the LHS and simplifying, can be rewritten as:

$$20 > x+\varepsilon - p_i \times (x+\varepsilon) + 2p_i \times (x+\varepsilon)$$

$$20 < x+\varepsilon+\varepsilon - p_i \times (x+\varepsilon+\varepsilon) + 2p_i \times (x+\varepsilon+\varepsilon)$$

Simplifying, we get:

$$20 > x + \varepsilon + p_i \times (x + \varepsilon)$$

$$20 < x + \varepsilon + \varepsilon + p_i \times (x + \varepsilon + \varepsilon)$$

Isolating p_i in the RHS, we get:

$$20 - (x + \varepsilon) > p_i \times (x + \varepsilon\ )$$

$$20 - (x + \varepsilon + \varepsilon) < p_i \times (x + \varepsilon + \varepsilon)$$

Which can be rewritten as:

$$\frac{20 - (x + \varepsilon)}{x + \varepsilon} > p_i$$

$$\frac{20 - (x + \varepsilon + \varepsilon)}{x + \varepsilon + \varepsilon} < p_i$$

Hence, p_i can be said to lie between the following boundaries:

$$\frac{20 - (x + \varepsilon + \varepsilon)}{x + \varepsilon + \varepsilon} < p_i < \frac{20 - (x + \varepsilon)}{x + \varepsilon}$$

QED.

8.2.2.2.2.2. More proofs on the social efficiency parameter elicitation

As we showed in corollary 5.1, the key value of the social efficiency parameter for our predictions of social efficiency preferences regarding cooperation attitudes in the SDG is $p_i \gtreqless \frac{2}{3}$. Below we provide a corollary showing that a maximum rejecting quantity (resp. minimum accepting quantity) of $x = 12$ reveals the necessary threshold for the social efficiency model to make predictions regarding play in the SDG.

Corollary 9.1. *Let's suppose that subject* i *maximizes the utility function* $U_i^{SE}(\pi_i,\pi_j)$*. Then,* if subject i's minimum accepting quantity is 12 or lower subject i reveals $p_i > \frac{2}{3}$. If subject i's maximum rejecting quantity is 12 or higher subject i reveals $p_i < \frac{2}{3}$

Given the inequality found in Proposition 9., it follows that a minimum accepting quantity of 12 or lower would entail:

$$p_i > \frac{20-(\leq 12)}{(\leq 12)}$$

Similarly, a maximum rejecting quantity of 2 or higher would entail:

$$p_i < \frac{20-(\geq 12)}{(\geq 12)}$$

And, hence,

$$p_i > \frac{\geq 8}{(\leq 12)}$$

$$p_i < \frac{\leq 8}{(\geq 12)}$$

Now, let's define a partially ordered set:

$$P:=(X,\leq)$$

Where

$$X \coloneqq \{x \in X | x \geq 0\}$$

We define the set MAQ as follows:

$$MAQ \coloneqq \left\{x \in MAQ| \left((x \in X) \wedge \left(x > \frac{\geq 8}{(\leq 12)}\right)\right)\right\}$$

And, also, we define the set MRO as follows:

$$MRQ \coloneqq \left\{x \in MRQ| \left((x \in X) \wedge \left(x < \frac{\leq 8}{(\geq 12)}\right)\right)\right\}$$

Using the same techniques as in in the previous corollaries., it is straightforward to see that $y = \frac{8}{12}$ is a lower bound of MAQ and an upper bound of MRQ. Hence, it follows that:

$$p > \frac{2}{3}$$

$$p < \frac{2}{3}$$

It follows that a person whose minimum accepting quantity is 12 or lower reveals $p_i > \frac{2}{3}$ and a person whose maximum rejecting quantity is 12 or higher reveals $p_i < \frac{2}{3}$.

QED.

8.2.2.2.3. Maximin parameter

8.2.2.2.3.1. Proof of proposition 10

Proposition 10. *Let's suppose that subject* i *maximizes the utility function* $U_i^{MM}(\pi_i, \pi_j)$.

(a) If subject i*'s maximum rejection quantity is* $x + \varepsilon$*, from the distribution* $(x + \varepsilon, x + \varepsilon)$*, to accept a distribution* $(20{,}0)$*, and if subject* i*'s minimum accepting quantity is* $x + \varepsilon + \varepsilon$*, from the distribution* $(x + \varepsilon + \varepsilon, x + \varepsilon + \varepsilon)$*, to reject a distribution* $(20{,}0)$*, then subject* i *will reveal to have a* q_i *parameter within the following boundaries:*

$$\frac{20 - (x + \varepsilon + \varepsilon)}{x + \varepsilon + \varepsilon} < q_i < \frac{20 - (x + \varepsilon)}{x + \varepsilon}$$

(b) If subject i*'s maximum rejection quantity is* $x + \varepsilon$*, from the distribution* $(x + \varepsilon, x + \varepsilon)$*, to accept a distribution* $(20{,}0)$*, and if subject* i*'s minimum accepting quantity is* $(x + \varepsilon + \varepsilon, x + \varepsilon + \varepsilon)$*, to reject a distribution* $(20{,}0)$*, then subject* i *reveals a maximin parameter* q_i *within the same threshold of values as the advantageous inequality parameter* β_i.

Proof.

(a)

Let's assume a person with a utility $U_i^{MM}(\pi_i, \pi_j)$ reveals the following preference pattern with their choices in the modified dictator games:

$$U_i^{MM}(20{,}0) > U_i^{MM}(x + \varepsilon, x + \varepsilon)$$

$$U_i^{MM}(20{,}0) < U_i^{MM}(x + \varepsilon + \varepsilon, x + \varepsilon + \varepsilon)$$

These equations can be rewritten as:

$$(1 - q_i) \times 20 + q_i \times (0) > x + \varepsilon$$

$$(1 - q_i) \times 20 + q_i \times (0) < x + \varepsilon + \varepsilon$$

Expanding the parenthesis, we get:

$$20 - q_i 20 > x + \varepsilon$$

$$20 - q_i 20 < x + \varepsilon + \varepsilon$$

Isolating p in the RHS, we get:

$$20 - (x + \varepsilon) > q_i \times 20$$

$$20 - (x + \varepsilon + \varepsilon) < q_i \times 20$$

Which can be rewritten as:

$$\frac{20 - (x + \varepsilon)}{20} > q_i$$

$$\frac{20 - (x + \varepsilon + \varepsilon)}{20} < q_i$$

Hence, q_i lies within the following boundaries:

$$\frac{20 - (x + \varepsilon + \varepsilon)}{20} < p < \frac{20 - (x + \varepsilon)}{20}$$

Which proves (a).

(b)

Recall the boundaries of β_i as found on proposition 8.:

$$\frac{20 - (x + \varepsilon + \varepsilon)}{20} < \beta_i < \frac{20 - (x + \varepsilon)}{20}$$

And recall the boundaries of q_i found in (a):

$$\frac{20-(x+\varepsilon+\varepsilon)}{20}<q_i<\frac{20-(x+\varepsilon)}{20}$$

Therefore, it follows that, given the generic maximum rejection quantity $x+\varepsilon$ and the minimum accepting quantity $x+\varepsilon+\varepsilon$, the boundaries of the maximin parameter q_i and of the advantageous inequality β_i will be the same, which proves (b).

QED.

8.2.2.2.3.2. More proofs on the maximin parameter elicitation

As we showed in corollary 6.1, the key value of the maximin parameter for our predictions of maximin preferences regarding cooperation attitudes in the SDG is $q_i \gtreqless 0.4$. Below we provide a corollary showing that a maximum rejecting quantity (resp. minimum accepting quantity) of $x = 12$ reveals the necessary threshold for the maximin model to make predictions regarding play in the SDG.

Corollary 10.1. *Let's suppose that subject i maximizes the utility function $U_i^{MM}(\pi_i, \pi_j)$. If subject i's minimum accepting quantity is 12 or lower, then subject i reveals $q_i > 0.4$. If subject i's maximum rejecting quantity is 12 or higher, then subject i reveals $p < 0.4$*

Proof.

Given that proposition 10. (b) shows that the values of β_i and q_i coincide for generic maximum rejection and minimum accepting quantities, this proof is identical to that of Corollary 8.1. (a) and, hence, has already been proven.

QED.

8.2.2.3. Reciprocity Games

We use a modified version of the reciprocity games used in Bruhin et al (2019) to elicit the $Y_{i,j}$ parameter values of the Dufwenberg and Kirchsteiger utility function outlined in chapter 4. We impose certain restrictions on the values of each of the three allocations strategically to simplify the finding on the threshold values for $Y_{i,j}$. More specifically, the allocations are such that some strategies are inefficient in Dufwenberg and Kirchsteiger's (2004) model, thereby simplifying the calculations. The paragraph below summarises our specific setting of the reciprocity games we present to subjects:

Person j could choose $a_j = E$, which will enforce the distribution (x_1, x_5), or alternatively could choose $a_j = N$, which would give person i the possibility to choose between $a_i = A$, generating a distribution of (x_2, x_4) and $a_i = B$, generating a distribution of (x_3, x_6), where $x_1 > x_2 > x_3$ and $x_4 > x_5 > x_6$.

It is important to note before proceeding that, given the Dufwenberg and Kirchsteiger (2004) model we use, the restrictions on the values we impose on x_1, x_2, x_3, x_4, x_5 and x_6 imply the following:

a) Strategy $a_i = B$ is inefficient, as $x_2 > x_3$ and $x_4 > x_6$, and hence both players would be better off by playing $a_i = B$.

b) Strategy $a_j = N$ is not inefficient. Whereas it is true that for one subsequent history of play (namely, $a_i = B$) both players end worse off by player j having played $a_j = N$, as $x_3 < x_1$ and $x_6 < x_5$, for at least another subsequent history of play ($namely, a_i = A$) at least one player is better off by player j having played $a_j = N$, as $x_4 > x_5$ even when $x_2 < x_1$.

8.2.2.3.1. Reciprocity parameter – Proof of proposition 11.

Proposition 11. *Let's suppose that subject i maximizes the utility function $U_i^{DK}\left(\pi_i, \pi_j\right)$. Then,*

(a) Assuming beliefs are in equilibrium, a player i's choice of $a_i = A$ over $a_i = B$ given that the first mover has done $a_j = N$ implies the following about the reciprocity parameter:

$$Y_{i,j} < \frac{2 \times (x_2 - x_3)}{(x_4 - x_6) \times (x_1 - x_2)}$$

(b) Assuming beliefs are in equilibrium, a player i's choice of $a_i = B$ over $a_i = A$ given that the first mover has done $a_j = N$ implies the following about the reciprocity parameter:

$$Y_{i,j} > \frac{2 \times (x_2 - x_3)}{(x_4 - x_6) \times (x_1 - x_3)}$$

Proof.

Given that the first mover has done $a_j = N$, the first-order belief of player i is updated so that $b_{ij}(h) = N$. The kindness function of player i towards player j reads:

$$\begin{aligned} &\kappa_i\left(a_{ij}(h), b_{ij}(h) = N\right) \\ &\qquad = \pi_j\left(a_{ij}(h), b_{ij}(h) = N\right) \\ &\qquad - \frac{max\, \pi_j\left(a_{ij}(h), N\right) | a_i \in A_i + min\, \pi_j\left(a_{ij}(h), N\right) | a_i \in E_i}{2} \end{aligned}$$

Hence, given that only $a_i = A$ is the only efficient strategy for player i as discussed above, it follows that:

$$\kappa_i\left(a_{ij}(h) = A, b_{ij}(h) = N\right) = x_4 - x_4 = 0$$

$$\kappa_i\big(a_{ij}(h) = B, b_{ij}(h) = N\big) = x_6 - x_4 = -(x_4 - x_6)$$

To find the perceived kindness function, note that $(p'', A; 1 - p'', B)$ is the probability distribution for the second-order belief of person i. Hence, we can write the perceived kindness function as:

$$\lambda_{iji}\left(b_{ij}(h) = N, c_{iji}(h)\right) = \pi_i\left(b_{ij}(h) = N, c_{iji}(h)\right) - \frac{x_1 + p'' \times x_2 + (1 - p'') \times x_3}{2}.$$

Using $(p'', A; 1 - p'', B)$ to compute the expected payoff that player j intends to give player i by doing $b_{ij}(h) = N$, we get:

$$\lambda_{iji}\left(b_{ij}(h) = N, c_{iji}(h)\right) = p'' \times \pi_i\big(b_{ij}(h) = N, a_i = A\big) + (1 - p'') \times \pi_i\big(b_{ij}(h) = N, a_i = B\big) - \frac{x_1 + p'' \times x_2 + (1 - p'') \times x_3}{2}.$$

Which, after substituting the relevant payoffs, becomes:

$$\lambda_{iji}\left(b_{ij}(h) = N, c_{iji}(h)\right) = p'' \times x_2 + (1 - p'') \times x_3 - \frac{x_1 + p'' \times x_2 + (1 - p'') \times x_3}{2}$$

Rearranging, we get:

$$\lambda_{iji}\left(b_{ij}(h) = N, c_{iji}(h)\right) = p'' \times x_2 + (1 - p'') \times x_3 - \frac{x_1}{2} - \frac{p'' \times x_2 + (1 - p'') \times x_3}{2}$$

Taking $p'' \times x_2 + (1 - p'') \times x_3$ as a common factor and simplifying, we get:

$$\lambda_{iji}\left(b_{ij}(h) = N, c_{iji}(h)\right) = -\frac{x_1}{2} + \frac{p'' \times x_2 + (1 - p'') \times x_3}{2} < 0$$

Given the perceived kindness that i believes j is displaying towards him, and the kindness of each possible action that i can do, we can write person i's utility of both actions as:

$$U_i\left(a_i(h) = A, b_{ij}(h) = N, c_{iji}(h)\right) = x_2 + Y_{i,j} \times (0) \times \left(-\frac{x_1}{2} + \frac{p'' \times x_2 + (1 - p'') \times x_3}{2}\right)$$

$$= x_2$$

$$U_i\left(a_i(h) = B, b_{ij}(h) = N, c_{iji}(h)\right)$$

$$= x_3 - Y_{i,j} \times (x_4 - x_6) \times \left(-\frac{x_1}{2} + \frac{p'' \times x_2 + (1 - p'') \times x_3}{2}\right)$$

(a)

For person i to choose the allocation which gives him the highest payoff ($a_i = A$) the following condition needs to hold:

$$U_i\left(a_i(h) = A, b_{ij}(h) = N, c_{iji}(h)\right) > U_i\left(a_i(h) = B, b_{ij}(h) = N, c_{iji}(h)\right)$$

Which is equivalent to the following expression:

$$x_2 > x_3 + Y_{ij} \times (x_4 - x_6) \times \left(\frac{x_1}{2} - \frac{p'' \times x_2 + (1 - p'') \times x_3}{2}\right)$$

Isolating $Y_{i,j}$ in the RHS, the previous expression becomes:

$$x_2 - x_3 > Y_{ij} \times (x_4 - x_6) \times \left(\frac{x_1}{2} - \frac{p'' \times x_2 + (1 - p'') \times x_3}{2}\right)$$

Dividing both sides of the inequality by $\left((x_4 - x_6) \times \left(\frac{x_1}{2} - \frac{p'' \times x_2 + (1 - p'') \times x_3}{2}\right)\right)$, we get:

$$Y_{i,j} < \frac{(x_2 - x_3)}{(x_4 - x_6) \times \left(\frac{x_1}{2} - \frac{p'' \times x_2 + (1 - p'') \times x_3}{2}\right)}$$

Let's assume that second-order beliefs are in equilibrium. That is to say, if $U_i\left(a_i(h) = A, b_{ij}(h) = N, c_{iji}(h)\right) > U_i\left(a_i(h) = B, b_{ij}(h) = N, c_{iji}(h)\right)$ then the second-order belief that Person i has is that Person j believes that he'll player $a_i(h) = A$ with certainty. Hence, $p'' = 1$. This would, in turn, give us the following threshold:

If $U_i\left(a_i(h) = A, b_{ij}(h) = N, c_{iji}(h)\right) > U_i\left(a_i(h) = B, b_{ij}(h) = N, c_{iji}(h)\right)$ and, hence, $p'' = 1$, then by substituting $p'' = 1$ in the inequality above, we get:

$$Y_{i,j} < \frac{2 \times (x_2 - x_3)}{(x_4 - x_6) \times (x_1 - x_2)}$$

(b)

If the beliefs are in equilibrium, it also follows that, if $U_i\left(a_i(h) = A, b_{ij}(h) = N, c_{iji}(h)\right) < U_i\left(a_i(h) = B, b_{ij}(h) = N, c_{iji}(h)\right)$, then the second-order belief that Person i has is that Person j believes that he'll play $a_i(h) = B$ with certainty. Hence, $p'' = 0$.

If $U_i\left(a_i(h) = A, b_{ij}(h) = N, c_{iji}(h)\right) < U_i\left(a_i(h) = B, b_{ij}(h) = N, c_{iji}(h)\right)$ and, hence, $p'' = 0$, then by substituting $p'' = 0$ in the inequality above, we get:

$$Y_{i,j} > \frac{2 \times (x_2 - x_3)}{(x_4 - x_6) \times (x_1 - x_3)}$$

QED.

9. Experimental instructions

Thank you for participating in our experiment.

In this experiment we will ask you to answer several questions. You will be paid a flat fee of £2.50 for completing this experiment. Additionally, provided you complete all elements of the experiment, you can win a bonus of up to £16.67 depending on your decisions and the decisions of other participants. We'll let you know which tasks may determine your bonus (and how) once you reach them.

Click >> to continue.

BEFORE YOU START!

1. Try to ensure that you will not be interrupted during the survey - close other applications and put other devices aside, so that you will not be distracted while completing the experiment. You will need to complete several tasks and it is important that you take them seriously.

2. Some general points on what to expect during the experiment:

- We will confront you with several decision situations and, in each of them, you will be paired at random with another participant.
- In each decision situation you can win points according to your and the other person's decisions.
- One of the decision situations will be picked at random.
- The one that is picked will be the one determining your payoff and the payoff of the person paired with you.
- The points you earned in the decision situation that is picked will be converted into pounds at the following rate: **Earnings in pounds = earnings in points / 6**
- In addition to completing those decision tasks, you must also answer some questions designed to gather some information about you and your views.
- We will wait until all participants have finished the experiments to make the pairs. Then, your payoff will be calculated and transferred to you.

Thank you.

Please, enter below your University of Nottingham email address and the email address to which your PayPal account is linked. We will use this information solely for the purposes of transferring your earnings from this experiment to your PayPal account. <u>Double check that you enter them correctly</u>, as otherwise we will not be able to process your payment!

Your PayPal account email address:

____

Your University of Nottingham email address:

____

[Each subject exposed to both the social dilemma game and the common interest game. *Different wording used for common interest game is introduced between brackets to avoid unnecessary repetition*]

Description of the Social Dilemma [Common Interest Game]

Please read the description below of the *'Group Project Dilemma'* decision problem

In this decision problem, Person A will interact with Person B.

Person A and Person B share a **group project**. Initially, there are 0 tokens in the project, but each person can contribute some tokens to it. Each person has control of 30 tokens and has four options: either contribute 0, 10, 20 or 30 tokens to the **group project**. Tokens someone does not contribute to the project are left in their **private account**.

Each person will receive an income from their private account and from the group project.

Income from their private account

Each person will receive 1 point for each token they leave in their private account. No one else receives anything from tokens that they leave in their own private account.

If, for example, Person A leaves 10 tokens in their private account, then Person A will receive 10 points from their private account and Person B will receive no points from Person A's private account.

Income from the group project

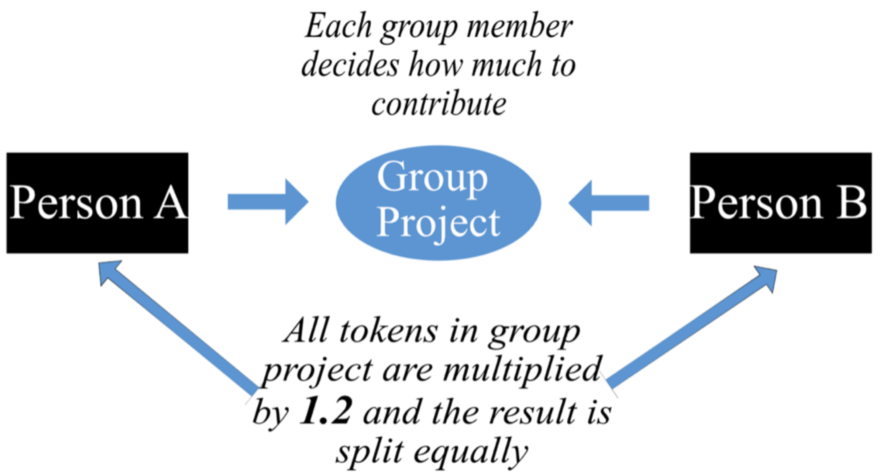
Each person benefits equally from tokens in the group project, regardless of who put them there. All tokens put in the project will be **multiplied by 1.2 [2.4], and the result will be split equally** among the two persons interacting.

If, for example, Person A contributes 10 tokens and Person B contributes 10 tokens to the project, then each of them will receive (10 + 10) × 1.2 [2.4] / 2 = 20 × 0.6 = 12 [24] points from the project.

Total income

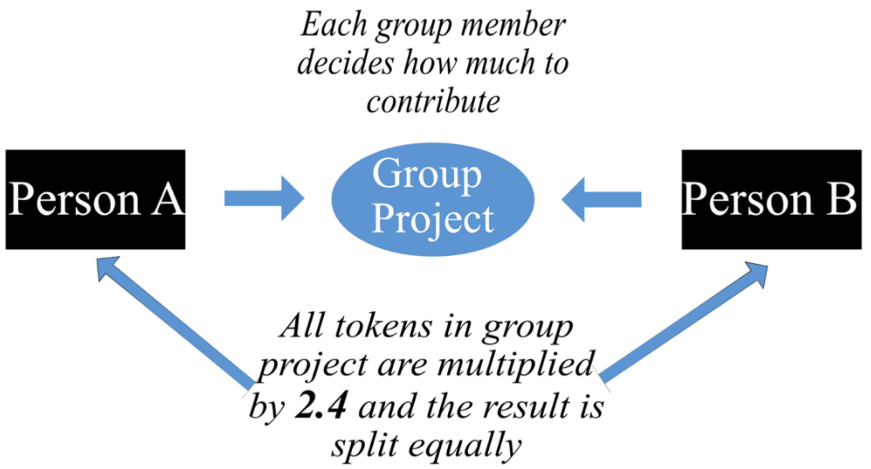
Each person receives the income from their own private account plus their share of income from the group project.

The figure below shows a summary of the interaction:


Each group member decides how much to contribute
Person A
Group Project
Person B
All tokens in group project are multiplied by **1.2** and the result is split equally

Each token in a person's private account earns that person 1 point
Income: point earnings from private account + point earnings from group project

(figure for the social dilemma game)


Each group member decides how much to contribute
Person A
Group Project
Person B
All tokens in group project are multiplied by **2.4** and the result is split equally

Each token in a person's private account earns that person 1 point
Income: point earnings from private account + point earnings from group project

(figure for the common interest game)

Please answer the following questions to check your understanding of the group decision problem.

Question 1.

Assume that Person A contributes 0 tokens to the group project and Person B contributes 0 tokens to the group project.

A) What will Person A's total point earnings be (total point earnings = point earnings from Person A's private account + point earnings from the group project)?

____

B) What will Person B's total point earnings be (total point earnings = point earnings from Person B's private account + point earnings from the group project)?

____

Question 2.

Assume that Person A contributes 30 tokens to the group project and Person B contributes 30 tokens to the group project.

A) What will Person A's total point earnings be (total point earnings = point earnings from Person A's private account + point earnings from the group project)?

____

B) What will Person B's total point earnings be (total point earnings = point earnings from Person B's private account + point earnings from the group project)?

____

Question 3.

Assume that Person A contributes 0 tokens to the group project and Person B contributes 30 tokens to the group project.

A) What will Person A's total point earnings be (total point earnings = point earnings from Person A's private account + point earnings from the group project)?

____

B) What will Person B's total point earnings be (total point earnings = point earnings from Person B's private account + point earnings from the group project)?

____

Question 4.

Assume that Person A contributes 20 tokens to the group project and Person B contributes 10 tokens to the group project.

A) What will Person A's total point earnings be (total point earnings = point earnings from Person A's private account + point earnings from the group project)?

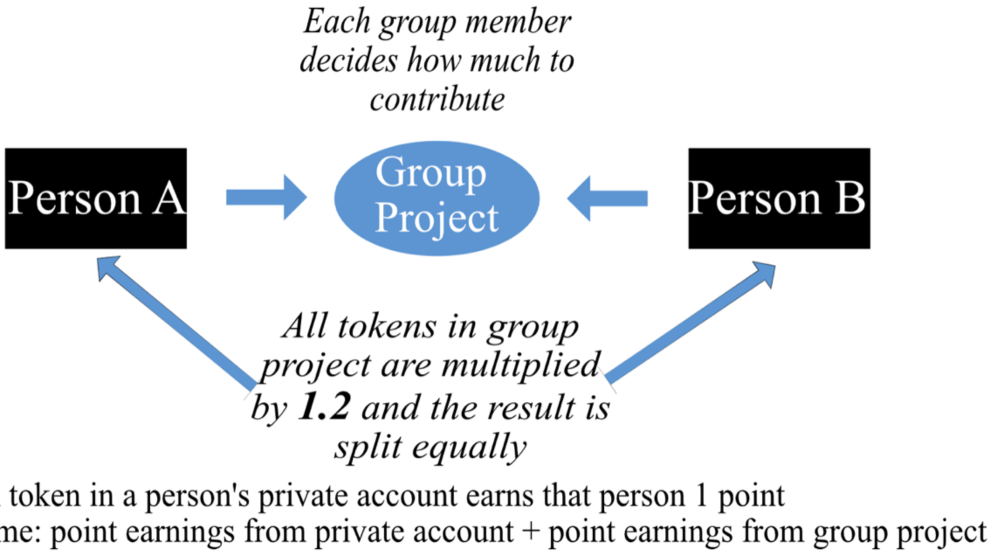
____

B) What will Person B's total point earnings be (total point earnings = point earnings from Person B's account + point earnings from the group project)?

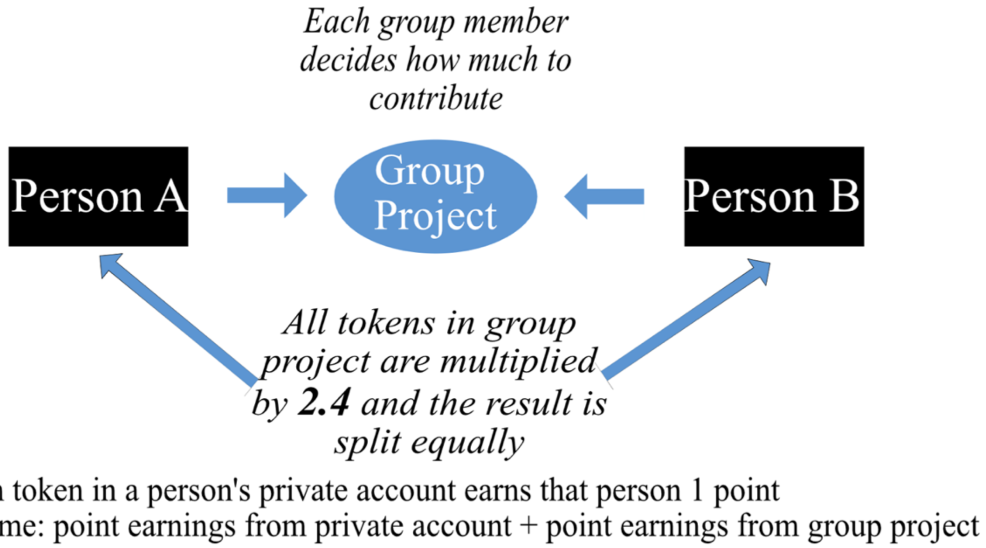
____

Instructions for the P-experiment

Your tasks here are based on the 'Group Project Dilemma' decision problem, which is summarised in the following figure:


Each group member decides how much to contribute
Person A
Group Project
Person B
All tokens in group project are multiplied by **1.2** and the result is split equally
Each token in a person's private account earns that person 1 point
Income: point earnings from private account + point earnings from group project

(figure for the social dilemma game)


Each group member decides how much to contribute
Person A
Group Project
Person B
All tokens in group project are multiplied by **2.4** and the result is split equally
Each token in a person's private account earns that person 1 point
Income: point earnings from private account + point earnings from group project

(figure for the common interest game)

In this decision situation, you interact with another person completing the experiment. You and the other person have two tasks, called the "**unconditional contribution**" and the "**contribution table**".

In the **unconditional contribution** task you simply decide the amount of tokens (either 0, 10, 20 or 30) you want to contribute to the group project.

In the **contribution table** task you indicate the amount of tokens **you want to contribute to** the group project **for each possible contribution of the other person**. Here, you can condition your contribution on that of the other person.

This is a one-off situation that is finished once you have made both decisions.

How your bonus from this decision situation, and the bonus of the other person you are paired with, will be determined (if this decision is chosen for payment)

The **unconditional contribution** task will be relevant for one of you and the **contribution table** task will be relevant for the other of you. Once you have finished the experiment, we will randomly decide which of you has the **unconditional contribution** task as relevant. If this decision situation is randomly chosen for payment, your choices in the relevant tasks will determine your payoffs as follows:

Example:

- The **unconditional contribution** task has been chosen to be relevant to Person A.
- Hence, Person B's **contribution table** will be relevant to Person B.
- Person A contributes 20 in the **unconditional contribution** task.
- In the **contribution table** task, Person B contributes 30 if Person A contributes 20.
- Hence, the total sum of contributions to the group project are 20 + 30 = 50 tokens.
- As a result, Person A earns 10 + 50 × 1.2 [2.4] /2 = 40 [72] points and Person B earns 0 + 50 × 1.2/2 = 30 [60] points.

Press continue when you are ready.

The unconditional contribution

How many tokens out of 30 do you contribute to the group project, i.e. 0, 10, 20 or 30?

____

The contribution table

Now we ask you to think about your contribution depending on how much the other person contributes. Please indicate for each possible contribution of the other person how much you contribute, i.e. 0, 10, 20 or 30.

	I contribute
If other contributes 0	
If other contributes 10	
If other contributes 20	
If other contributes 30	

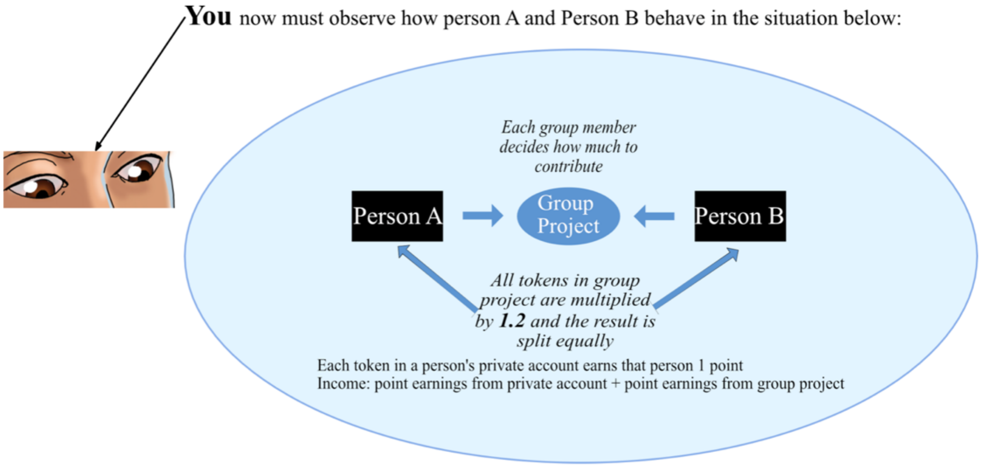
Instructions for the M-experiment

The goal of the following tasks is to investigate **your own** moral views of the **'Group Project Dilemma'** decision problem. These tasks will be presented in the next screens.

There are no correct or incorrect answers - just respond with what **you really think**

Press continue when you are ready.

You are now an outside OBSERVER of the 'Group Project Dilemma' decision problem described earlier and summarized in the following picture.



You now must observe how person A and Person B behave in the situation below:

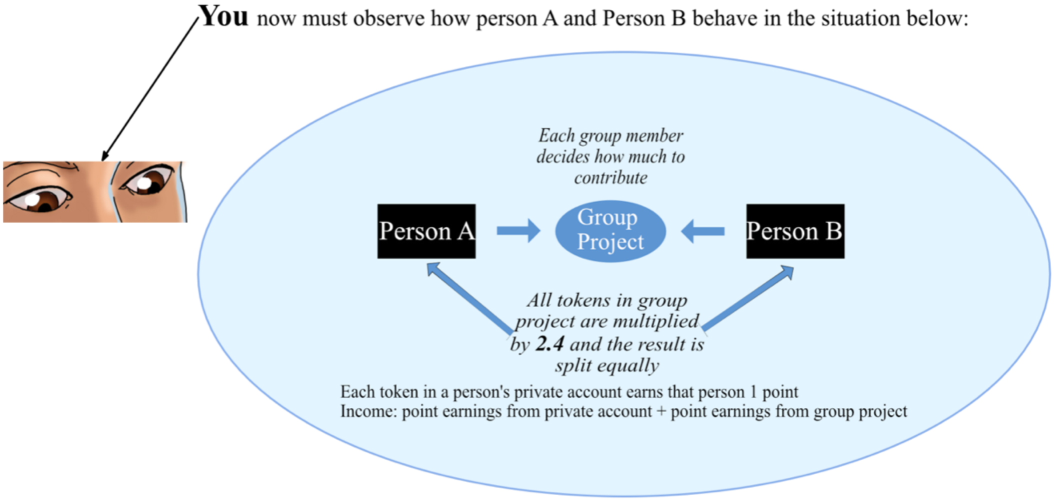
Each group member decides how much to contribute

*All tokens in group project are multiplied by **1.2** and the result is split equally*

Each token in a person's private account earns that person 1 point

Income: point earnings from private account + point earnings from group project

(figure for the social dilemma game)



You now must observe how person A and Person B behave in the situation below:

Each group member decides how much to contribute

*All tokens in group project are multiplied by **2.4** and the result is split equally*

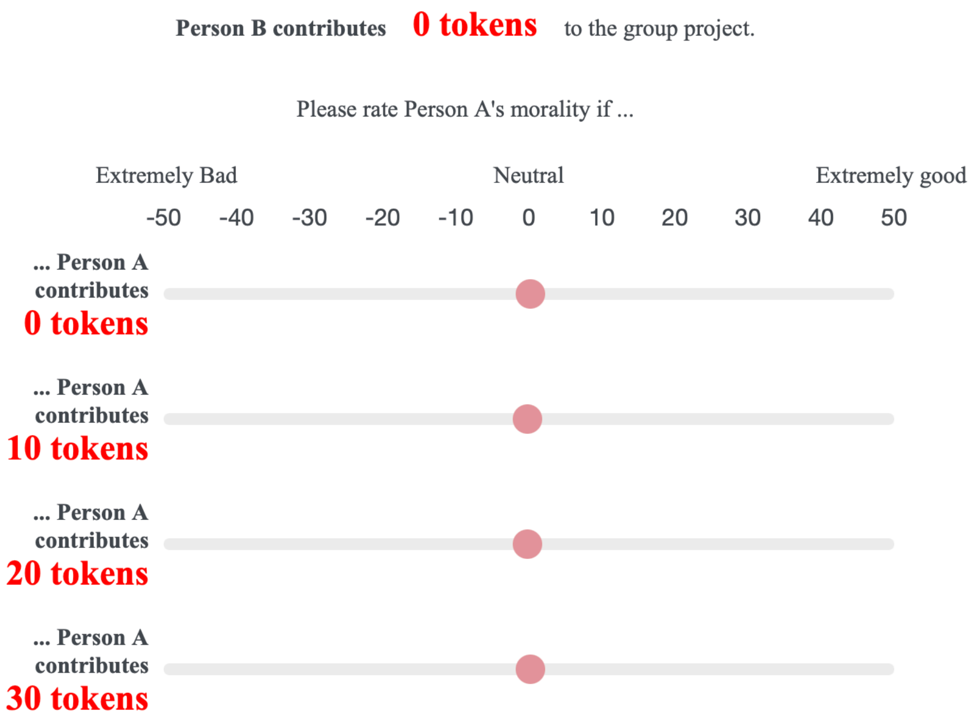
Each token in a person's private account earns that person 1 point

Income: point earnings from private account + point earnings from group project

(figure for the common interest game)

Your task as an observer is to give your moral rating of Person A in scenarios that we'll present you in the following screens.

Rate the morality of Person A on a scale from -50 (extremely bad) to +50 (extremely good) with the sliders provided. In each case you must click on the slider to activate it and then move it to the rating you decide on.



Person B contributes **0 tokens** to the group project.

Please rate Person A's morality if ...

Extremely Bad — Neutral — Extremely good

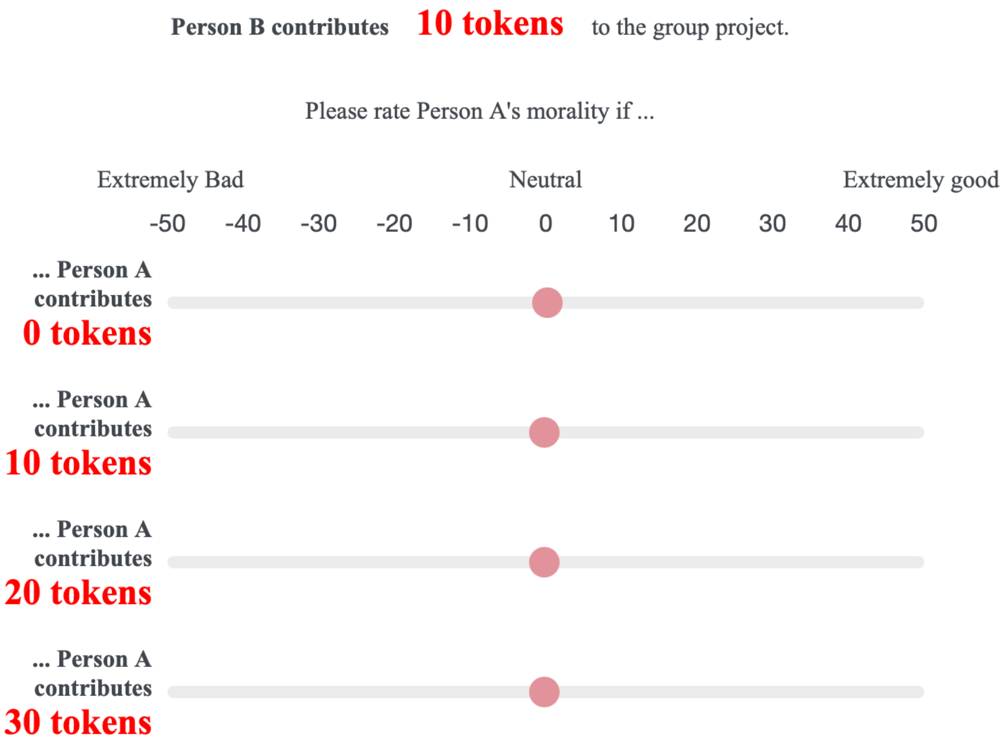
-50 -40 -30 -20 -10 0 10 20 30 40 50

... Person A contributes **0 tokens**

... Person A contributes **10 tokens**

... Person A contributes **20 tokens**

... Person A contributes **30 tokens**



Person B contributes **10 tokens** to the group project.

Please rate Person A's morality if ...

Extremely Bad — Neutral — Extremely good

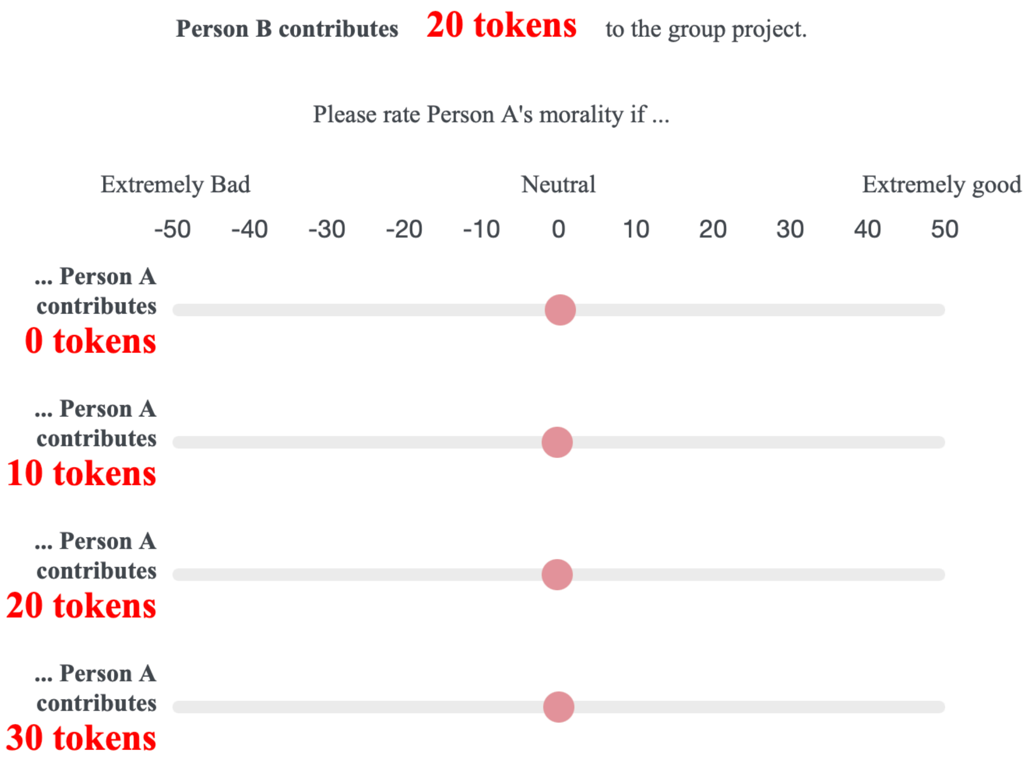
-50 -40 -30 -20 -10 0 10 20 30 40 50

... Person A contributes **0 tokens**

... Person A contributes **10 tokens**

... Person A contributes **20 tokens**

... Person A contributes **30 tokens**



Person B contributes **20 tokens** to the group project.

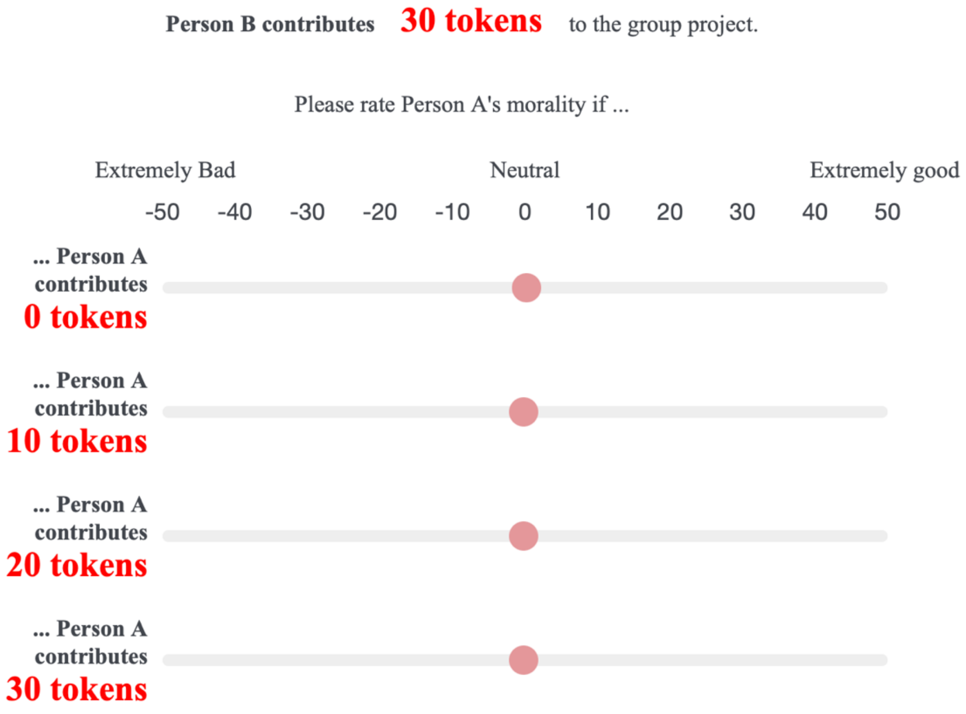
Please rate Person A's morality if ...

... Person A contributes **0 tokens**

... Person A contributes **10 tokens**

... Person A contributes **20 tokens**

... Person A contributes **30 tokens**



Person B contributes **30 tokens** to the group project.

Please rate Person A's morality if ...

... Person A contributes **0 tokens**

... Person A contributes **10 tokens**

... Person A contributes **20 tokens**

... Person A contributes **30 tokens**

Instructions for the parameter-elicitation games

Instructions for the Ultimatum Game

Please read the description below of the *'proposal'* decision problem

In this decision problem, a ***proposer*** will interact with a ***responder***. The decision problem is as follows:

- The proposer's decision is to propose a distribution of a fixed number of points between themself and the responder.
- The responder can accept or reject the proposer's distribution.
- If the responder accepts, the proposer's distribution will determine the points each gets.
- If the responder rejects, both receive 0 points.

Press continue when you are ready.

Ultimatum Game: decision-making clarification

You are now taking part in a decision situation based on the '*proposal*' decision problem

- You will have two different tasks
- In the '*proposer task*', you will decide the distribution you want to propose to the responder
- In the '*responder task*', you will decide whether to accept or reject each proposal that the proposer could have made.
- One task will be relevant for one of you and the other task will be relevant for the other of you. Once you have finished the experiment, we will choose who of you has the '*proposer task*' as relevant. If this decision situation is randomly chosen for payment, your choices in the relevant tasks will determine your payoff and that of the participant you are paired with.

Press continue when you are ready.

Proposer task

Which of the following distributions do you want to propose to the responder?

- 14 points for me, 0 points for the responder
- 13 points for me, 1 point for the responder
- 12 points for me, 2 points for the responder
- 11 points for me, 3 points for the responder
- 10 points for me, 4 points for the responder
- 9 points for me, 5 points for the responder
- 8 points for me, 6 points for the responder
- 7 points for me, 7 points for the responder

Responder task

Will you accept or reject each of the following proposals if they were made by the proposer?

Choose Accept if you want to accept a given proposal and Reject otherwise

	Accept	Reject
14 points for the proposer, 0 points for me		
13 points for the proposer, 1 point for me		
12 points for the proposer, 2 points for me		
11 points for the proposer, 3 points for me		
10 points for the proposer, 4 points for me		
9 points for the proposer, 5 points for me		
8 points for the proposer, 6 points for me		
7 points for the proposer, 7 points for me		

Instructions for the Reciprocity Games

Please read the description below of the *'delegation'* decision problem

In this decision problem, the ***first mover*** will interact with the ***second mover***. The decision problem is as follows:

- The first mover has to choose between selecting a ***Default Distribution*** or delegating to the second mover the decision of selecting between ***Distribution A*** and ***Distribution B***.
- The ***Default Distribution***, ***Distribution A*** and ***Distribution B*** are alternative distributions of points between the first mover and the second mover.
- If the first mover selects the ***Default Distribution***, then that distribution will determine the points of each of them. If the first mover delegates to the second mover the decision of selecting between ***Distribution A*** and ***Distribution B***, then the distribution that the second mover selects will determine the points of each of them

Press continue when you are ready.

Reciprocity Games: decision-making clarification

You are now taking part in several decision situations based on the 'Delegation' decision problem.

- You will have two different tasks.
- In the '*first mover tasks*', you will choose, for each decision situation, between selecting the ***Default Distribution*** or delegating to the second mover the decision of selecting between ***Distribution A*** and ***Distribution B***.
- In the '*second mover tasks*', you will act, in each decision situation, as if the first mover had delegated the decision of selecting between Distribution A and Distribution B to you. That is, you will select one of either distributions.

<u>How you bonus from this decision situations, and the bonus of the person you are paired with, will be determined</u>

- Once you have finished the experiment, we will choose who of you has the '*first mover tasks*' as relevant. And, also, which of all the decision situations will be relevant for both of you.
- For the relevant decision situation, if the person having the first mover tasks as relevant chooses the ***Default Distribution***, then the ***Default Distribution*** will determine your payoffs.
- For the relevant decision situation, if the person having the first mover tasks as relevant chooses delegating, then the choice of the other person in the second mover tasks will be relevant for payment. And, your payoffs will be determined by the Distribution that this other person chooses (either ***Distribution A*** or ***Distribution B***

Press continue when you are ready

First mover tasks

The ***Default Distribution*** and ***Distribution A*** are **the same in all decision situations**, but ***Distribution B*** varies accross **decision situations**.

The ***Default Distribution*** and ***Distribution A*** for all the decision situations are shown at the top of the table. **Each row of the table represents a decision situation**, and ***Distribution B*** for a given decision situation is provided at the left of each row.

RG_First_Choice **Do you want to select the *Default Distribution* or delegate to the second mover the decision of selecting between *Distribution A* and *Distribution B*?**

The *Default Distribution* and *Distribution A* are:

***Default Distribution*: 5** points for **me, 95** points for the **second mover**
***Distribution A:* 0** points for **me, 0** points for the **second mover**

	Select ***Default Distribution***	Delegate to the second mover
Distribution B: **100** points for **me**, **0** points for the **second mover**		
Distribution B: **85** points for **me**, **15** points for the **second mover**		
Distribution B: **81** points for **me**, **19** points for the **second mover**		
Distribution B: **80** points for **me**, **20** points for the **second mover**		
Distribution B: **75** points for **me**, **25** points for the **second mover**		
Distribution B: **70** points for **me**, **30** points for the **second mover**		
Distribution B: **60** points for **me**, **40** points for the **second mover**		
Distribution B: **43** points for **me**, **57** points for the **second mover**		
Distribution B: **29** points for **me**, **71** points for the **second mover**		
Distribution B: **22** points for **me**, **78** points for the **second mover**		
Distribution B: **8** points for **me**, **92** points for the **second mover**		

Second mover tasks

The ***Default Distribution*** and ***Distribution A*** are **the same in all decision situations**, but ***Distribution B*** varies accross **decision situations**.

The ***Default Distribution*** and ***Distribution A*** for all the decision situations are shown at the top of the table. **Each row of the table represents a decision situation**, and ***Distribution B*** for a given decision situation is provided at the left of each row.

If the first mover were to delegate the decision of selecting between *Distribution A* and *Distribution B*, which of them would you choose in each decision situation?

The *Default Distribution* and *Distribution A* are:

***Default Distribution*: 5** points for the **first mover, 95** points for **me**

***Distribution A*:0** points for the **first mover, 0** points for **me**

	Select ***Distribution A***	Select ***Distribution B***
***Distribution B*: 100** points for the **first mover**, **0** points for **me**		
***Distribution B*: 85** points for the **first mover**, **15** points for **me**		
***Distribution B*: 81** points for the **first mover**, **19** points for **me**		
***Distribution B*: 80** points for the **first mover**, **20** points for **me**		
***Distribution B*: 75** points for the **first mover**, **25** points for **me**		
***Distribution B*: 70** points for the **first mover**, **30** points for **me**		
***Distribution B*: 60** points for the **first mover**, **40** points for **me**		
***Distribution B*: 43** points for the **first mover**, **57** points for **me**		
***Distribution B*: 29** points for the **first mover**, **71** points for **me**		
***Distribution B*: 22** points for the **first mover**, **78** points for **me**		
***Distribution B*: 8** points for the **first mover**, **92** points for **me**		

Instructions for the Modified Dictator Games

Please read the description below of the *'no-rejection'* decision problem

In this decision problem, the ***first mover*** will interact with the ***passive person***. The decision problem is as follows:

- The first mover has to choose between two different distributions of points between themself and the passive person.
- The passive person has no choice but to accept what the first mover chooses.
- Points each of them gets are determined by the first mover's chosen distribution Press continue when you are ready.

Modified Dictator Games: decision-making clarification

You are now taking part in several decision situations based on the '*no-rejection*' decision problem.

- You will choose between the two distributions of points available.
- If this decision problem is chosen for payment, only one of the decision situations will be chosen at random for payment.
- Once you have finished the experiment, we will choose who of you has the tasks as relevant and who acts as the passive person. If this decision problem is randomly chosen for payment, your choice (if you are chosen to act as the first mover) in the chosen decision situation will determine your payoffs.

Press continue when you are ready

Dictator tasks

You can choose *Distribution 1* or *Distribution 2*, where *Distribution 2* is the same in all decision situations. *Distribution 1* is different in all decision situations.

Do you want to choose Distribution 1 or Distribution 2?

***Distribution 2*: 20** points for **me, 0** points for the **passive person**

	Choose ***Distribution 1***	Choose ***Distribution 2***
***Distribution 1*: 0** points for **me**, **0** points for the **passive person**		
***Distribution 1*: 2** points for **me**, **2** points for the **passive person**		
***Distribution 1*: 4** points for **me**, **4** points for the **passive person**		
***Distribution 1*: 6** points for **me**, **6** points for the **passive person**		
***Distribution 1*: 8** points for **me**, **8** points for the **passive person**		
***Distribution 1*: 10** points for **me**, **10** points for the **passive person**		
***Distribution 1*: 12** points for **me**, **12** points for the **passive person**		
***Distribution 1*: 14** points for **me**, **14** points for the **passive person**		
***Distribution 1*: 16** points for **me**, **16** points for the **passive person**		
***Distribution 1*: 18** points for **me**, **18** points for the **passive person**		
***Distribution 1*: 20** points for **me**, **20** points for the **passive person**		
***Distribution 1*: 22** points for **me**, **22** points for the **passive person**		
***Distribution 1*: 24** points for **me**, **24** points for the **passive person**		
***Distribution 1*: 26** points for **me**, **26** points for the **passive person**		
***Distribution 1*: 28** points for **me**, **28** points for the **passive person**		
***Distribution 1*: 30** points for **me**, **30** points for the **passive person**		
***Distribution 1*: 32** points for **me**, **32** points for the **passive person**		

Sociodemographics Questionnaire

[Each sentence was displayed with Font Times New Roman, size 18, bold and left-aligned. Unless otherwise stated, The options for the respondent in each question of the sociodemographic questionnaire appeared on a dropdown list below each of the statements. We provide the options for each questions below the question itself]

Q1. Your Gender:
[Options to the respondent: Male, Female, Prefer not to say]

Q2. Your Age:
[Options to the respondent: from 15 to 100 in steps of 1]

Q3. Would you describe yourself as a left wing or a right wing?

[Options to the respondent: Neutral, Left, Very Left, Right, Very Right,, Prefer not to say]

Q4. How religious are you?

[Options to the respondent: Not at all, Somewhat religious, Very religious, Prefer not to say]

Q5. How large was the community where you have lived the most time of your life?

[Options to the respondent: Up to 2,000 inhabitants, Between 2,000 and 10,000 inhabitants, Between 10,000 and 100,000 inhabitants, More than 100,000 inhabitants]

Q6. What is your field of study?

[The question was open-ended: students introduced their subject directly]

Q7. Here are a number of personality traits that may or may not apply to you. Please indicate on the scale below the extent to which you agree or disagree with that statement. You should rate the extent to which the pair of traits applies to you, even if one characteristic applies more strongly than the other.

Extraverted, enthusiastic
Critical, quarrelsome
Dependable, self-disciplined
Anxious, easily upset
Open to new experiences, complex
Reserved, quiet
Sympathetic, warm
Disorganised, careless
Calm, emotionally stable
Conventional, uncreative

[Options to the respondent: Disagree strongly, Disagree moderately, Disagree a little, Neither agree nor disagree, Agree a little, Agree moderately, Agree strongly]

[This question was presented in a matrix table, with the personality traits in the y-axis and the options to the respondent in the x axis]

Last question before leaving

Which, if any, of the following concepts were you taking into account when making your choices in the decision problems we have presented to you earlier? Select as many as apply to you

Notes: You may have some doubts as to which option(s) to choose, as many of the different concepts we present were relevant for the decision situation. Below we provide you with two points to help you better assess your answer to the question.

It may happen that two or more concepts were relevant for your understanding of the decision problem, but that only one of those was the reason underlying your choices. In this case, you should choose only the concept that was the reason for your choice. It may happen that many concepts were underlying your choices, either because (i) you were taking into account different concepts for making your choices in different decision problems, or (ii) because you cared about different concepts when making your choices. If either (i) or (ii) apply to you, please choose all the concepts underlying your choices.

- Avoid inequality
- Be reciprocal
- Avoid doing what I consider to be morally bad
- Do what I consider the most morally good
- Increasing my own payoff
- Increasing the payoff of the other person paired with me
- Increasing the payoff of the person getting the lowest payoff from the interaction
- Increasing the total payoff that I and the person paired with me get
- Maximise my own happiness, regardless of how broadly my happiness is defined to be (e.g. your happiness can depend solely on your own payoff, but it can also be influenced by any concept that you can think of, such as the level of inequality that derives from your choice, by how morally good the action you think about doing is, etc).
- Other. Please, specify

____