FROM

# MORALITY

TO

# Rules to Choices:

## INTRODUCING and TESTING a New Theory

ON HOW

# MORALS

# INFLUENCE COOPERATION.

> *"... there are mental passions, by which we are impelled immediately to seek particular objects ... without any regard to interest; and when these objects are attained, a pleasuring enjoyment ensues, as the consequence of our indulged affections"*
>
> DAVID HUME

> *"... and perhaps the reasons to be offer'd may prove ... that what excites us to these Actions which we call virtuous, is ... an entirely different Principle of Action from Interest or Self-Love"*
>
> FRANCIS HUTCHESON

*NOTTINGHAM:*

Printed for CHRIS STARMER & BERTIL TUNGODDEN, at the

*University of Nottingham.*

MMXXII.

# From Morality to Rules to Choices:

## Introducing and Testing a New Theory on how Morals influence Cooperation

**Ernesto María Gavassa Pérez**

*Supervisors:*

**Robin Patrick Cubitt**

**Simon Gächter**

**Centre for Decision Research and Experimental Economics (CeDEx)**

**School of Economics**

**University of Nottingham**

**This dissertation is submitted for the degree of Doctor of Philosophy**

**February 2022**

*<<The Fell Types used in the cover and as headers of the thesis are digitally reproduced by Igino Marini. www.iginomarini.com>>*

*(this page intentionally left blank)*

# DECLARATION

This dissertation is the result of my own work, and what is included within represents my original and individual work undertaken during my PhD. Chapters 2 and 3 will be used to develop future co-authorship papers with my two supervisors, but I assert that all the work regarding those chapters, as well as all the others, has been solely produced by me, and has immensely benefitted from the guidance and feedback naturally ensuing from a supervision process.

In accordance with the University guidelines, and with that of common sense and good academic practice, I assert that none of the content within the thesis has been plagiarised; and that I have cited, where appropriate and to the best of my abilities, all the sources that led to the development of the novel ideas presented herein.

I declare that the version of the thesis contained in this document is the final version I submit for examination and has been approved by both my supervisors via the intention to submit document, and via several readings of the chapters after signing the document so that we agreed on a content that was deemed to be of good quality for the exam.

Signed:

,

Date: 27.02.2022

# Dedicatory

*To you, Mum.*

*In the nights*
*Where I was scared*
*I used to cry,*
*And call your name.*
*You always came,*
*Erased the shadows,*
*Gave me strength*
*And changed my fate.*

*Now I am grown,*
*I cry no more.*
*But without you*
*I don't have soul.*
*Through darker times*
*And wuthering storms,*
*I feel your presence:*
*I feel your love.*

*And, to the memory of my late father.*

*(this page intentionally left blank)*

*(this page intentionally left blank)*

# Acknowledgments

In the acknowledgments section of my thesis, I want to be fair, and recognise everyone that has contributed to my intellectual development, and personal joy, not only during my PhD years but through all my life. It is a hard endeavour, not only because I will surely forget some names, but because some people will feel disappointed in the order in which I present, and thank, everyone, as they may have wanted to be somewhat up in the list. In no way being lower in the list means a lower importance, as all I can feel is an enormous gratitude to everyone that formed me as a human being and that is included in the lines that follow.

It will be impossible to fill this list without putting my mother above everyone else, by a huge distance. She is, and means, everything to me. Since I was a child, but especially since my father died, and even when she was very ill, she never omitted a smile, a caress, an act of love. Without you, I would certainly not have been able to pursue an academic career that I so much love; and, for that, and many other things, you make the first place in my list. You will always be the most important human being to ever have passed through my life: I love you beyond necessity.

It seems strange to me how can men that I do not even know, and that I will never have the pleasure to know, have influenced me so much, and given me so much joy during my PhD. When I started this journey, back in October 2017. Well, actually, let's stop here for a minute and go back a few months. When I was doing my master in Nottingham, Daniele Nosenzo, to whom I admire and owe my passion towards behavioural economics, taught me the amazing world of prosocial preferences. I have never seen anything particularly exciting, or that could account to any serious knowledge, in the neoclassical benchmark of a selfish maximiser, regardless of what a constructivist, an anti-realist, or a rationalist (in the epistemological sense of the three terms) may argue. And then, magic happened. This Italian guy was teaching Fehr-Schmidt, and I felt elated: this is what I have been thinking all my life! I was raising my hand every five minutes to ask questions, and there is where I knew that I could never derive any pleasure in the benchmarks imposed by neoclassical thinking. We could say with some certainty, if certainty is to exist regarding one's own feelings, that when I started my PhD, I was a fan of prosocial preferences. And, when a dramatic event happened in my life I don't know why but I started to read David Hume. A book called '*An Enquiry Concerning the Principles of Morals*'. There, my life changed. Forever. It is, without a doubt, the most penetrating essay regarding prosociality that I have ever read, although I venture that (and I hope I am wrong, because it goes against my own research), economists will take a long time to accept, if they end up doing, the radicality of his arguments and their implications. It is the first, and only man, except for Imre Lakatos, that has made me cry of joy, with his sophisticated, delicate, and profound writing. He challenged all my conceptions about prosociality that Daniele Nosenzo had seeded in me. ALL. None escaped. Appendix II, *Of Self-love*, is the briefest and most powerful argument ever to be written against prosocial preferences. I left the book because I was feeling disturbed. But then, one year later, I opened it again; and that disturbance, that challenge, transformed slowly and naturally into love. So, this thesis is, beyond the obvious dedication to my mother, a dedication and homage to that privileged mind, and to his worthy, and promising, theory of human's moral nature.

David Hume guided me to Thomas Hobbes, Francis Hutcheson, Adam Smith, and the Earl of Shaftesbury. It is to those, as well, that I want to acknowledge their contribution to the development of my views, that I have tried to use as an inspiration for the MRC framework. Hobbes inspired blame avoidance, and Shaftesbury, Hutcheson, and Hume inspired praise seeking alongside the main structure of the MRC framework, to which I also owe, in the greatest degree, the influence of Vernon Smith and Bart Wilson's works.

Now, let's go back to the 21st century. There have been a lot of people that have helped me disinterestedly. That is, they did never expect, nor did they ask, anything from me, and, yet they gave me all the help they could. It is to those people that I feel the most gratitude, and the greatest admiration for such a beautiful virtue in their character. To all my colleagues in Barcelona, and Madrid, that once wrote me reference letters, I want to acknowledge. To Elke Renner, Alex Possajennikov, and Silvia Sonderegger, who also wrote me reference letters, I want to acknowledge. To Alex also for his help and advice on the third chapter, the greatest of my respects and acknowledgments, and to Martin Sefton, for his comments. To Chris Starmer, for his support during the master and for having his door open every time I needed it, I also want to acknowledge.

To Vernon Smith, that discussed with me about the MRC framework and gave me advice on how to frame it in two occasions, I also acknowledge to the greatest degree.

To Bertil Tungodden, who also gave me advice on the MRC framework, and that additionally gave me several opportunities without me asking for them, I also acknowledge to the greatest of the degrees. To Bob Sugden, thanks for our chat about Hobbes, and for all the advice you have given me so far. To Urs Fischbacher, for our long talk when he came to Nottingham and for the hospitality he showed when I went to the z-Tree course in Konstanz. To all the other speakers that came to Nottingham, I also acknowledge, especially to Colin Camerer, Michalis Drouvelis, and Theo Offerman, who gave me extensive feedback and amazing company.

What can I say about my supervisors, and co-authors of the work that will come out of chapters 2 and 3? Of course, an appreciation of a greater degree ensues for their feedback, constant appraisal and encouragement, and care. Robin Cubitt, thanks for all your help and generosity these years, and for the very detailed, precise, and gentle feedback you gave me of the versions of the chapters included herein. Simon Gächter, I can never be grateful enough for the honest, intellectual environment that you provided me: you were always open, and received well, my critiques to prosocial preferences once Hume entered into my mind, and you always were critical enough to assess and reflect on my points of view, and develop a great respect towards me for my intellectual development throughout the years, even when our ideas on some scientific topics collided (most of them did not). I can only but feel the most gratitude to you, and to our friendship: to our coffees together, to our lunches, to our laughs, and to the mutual understanding and respect we built during these years. I sincerely hope your new grant gave you what you wanted and wish you the best in the years to come: you shall achieve whatever is you were looking for.

To my various friends throughout these years in Nottingham; that is, to Nina Serdarevic, Wenxiao, Abubakar, Adrian, Gabriele (Atletico strictly better than Juve, if in doubt ask Fabio), to Leo, to Martina, to Jesal, I want to acknowledge as well. To all the friends I made in Norwich and Amsterdam (that is, Kevin, Konstantinos, Mike, Prachi, and Rosie, among others), and to all the CeDEx team for such a wonderful environment, I want to acknowledge. To Anna Hochleitner, Monica Pompeo, Nina Serdarevic, Konstantinos Ioannidis, and Michael Brock for giving feedback on previous versions of some chapters, also an acknowledgment is due. Finally, to everyone that once helped me, my most sincere gratitude and acknowledgment ensues.

*Ernesto M. Gavassa Pérez*

*(this page intentionally left blank)*

# Abstract

This thesis presents a framework that I name the MRC framework, as its purpose is to capture how people go from *Morality* to *Rules* to *Choices*. I embed two theories, Blame Avoidance and Praise Seeking, within the MRC framework. The former states that subjects are impelled to avoid what they consider as blameworthy strategies from an impartial perspective, given all the circumstances that surround their evaluation of such strategies. The latter states that subjects are impelled to seek doing what they consider as the most praiseworthy strategies from an impartial perspective, given all the circumstances that surround their evaluation of such strategies. The chief characteristic feature that distinguishes my new models from those present in the literature is that they explore the motivational force of prosociality from an impartial perspective, thereby replacing the characteristic self-centredness of most models of other-regarding preferences within the literature. My theories and classical models of other-regarding preferences need not be either orthogonal or theories making different predictions. Rather, what I contend is that their ultimate explanation for the flourishing of prosocial behaviour is radically different. Whereas social preferences contend that subjects are moved by their self-interest, howsoever enlightened and altruistically inclined, our theories propose that self-interest plays a minor role, if any at all, in prosocial considerations.

As a first step towards this end, the thesis focuses on studying how these newly developed theories fare at explaining behaviour at public goods games, a canonical form of a social dilemma. The thesis contains three core chapters – chapter 2, chapter 3, and chapter 4 – plus an introduction (chapter 1) and a conclusion (chapter 5). The first of those (chapter 2) starts by developing an elicitation tool that allows us to measure empirically the moral judgments of a person, from an impartial perspective. Chapter 3 uses the newly developed tool and the theories proposed to test the explanatory of the MRC framework in predicting unconditional contributions and contribution attitudes to give-some and take-some public goods. Chapter 4 goes beyond the scope of chapter 3 and makes a direct test of the MRC framework against several canonical models of social preferences at the individual level: material selfishness, inequality aversion, reciprocity, social efficiency, maximin, and spite. The test explores the explanatory power of each of the theories at predicting contribution

attitudes of a social dilemma (MPCR <1) and a common interest game (MPCR >1). Our results show that (i) social dilemmas are perceived as moral issues (chapter 2); (ii) blame avoidance can predict contribution attitudes of both give and take social dilemmas, and blame avoidance and praise seeking can predict unconditional contributions in give and take social dilemmas (chapter 3); and (iii) that blame avoidance, along with inequality aversion and maximin, is among the three best performing theories in predicting contribution attitudes to social dilemmas and common interest games (chapter 4).

# THE CONTENTS.

*(this page intentionally left blank)*

# List of Tables

*(this page intentionally left blank)*

# List of Figures

# Chapter 1. Introduction

The purpose of this thesis is to enquire whether in a society, in its natural state, that is, with no enforceable norms, and no views to future rewards or punishment, or reputational concerns confounding motivations for action, people still perform cooperative actions; and to propose two different mechanisms for the flourishing of such actions: that of *blame avoidance,* viz., a normative determination to avoid performing that which one perceives as blameworthy, and that of *praise seeking*, viz., a normative determination to seek performing that which one perceives as most praiseworthy, from an impartial standpoint, and given the circumstances that surround their normative evaluations of all potential actions available at the moment of choice. The two theories I propose in this thesis, blame avoidance and praise seeking, provide workable models that capture how people's moral judgments influence cooperation.

Morality plays an important part in our daily lives. According to Hofmann et al (2014), who surveyed several people in distinct times of their day, subjects self-reported that around 29% of their daily live situations were of moral significance. Yet, to date, and as Ellemers et al (2019) suggest after surveying the content of 1,278 papers published in moral and social psychology between 1940 and 2017, there is very little work done with regards to the relation between moral judgments and behaviour in decision situations. In their own words, '*substantial knowledge has accumulated about the way people think about morality; however, we know much less about how this affects their moral behaviour*' (pp.354). The motivation of this thesis is to explore whether the newly developed theories of blame avoidance and praise seeking influence people's behaviour in cooperation problems, thereby allowing us to contribute to advancing the knowledge within the aforementioned gap.

The relation between ethics and economics has been the source of an important debate between economists, and there is no homogeneous view on the subject. On the one hand, some economists defend the position that ethics and economics influence each other. As an example of a defence of this view, see Alfred Marshall's claim:

> "*Ethical forces are among those of which the economist has to take account*"
>
> Alfred Marshall (2013), *Principles of Economics*

On the other hand, other economists contend that normative judgments should play no role in what we know as positive economics. As an example of this view, see Milton Friedman's claim:

> "*Positive economics is in principle independent of any particular ethical position or normative judgments.*"
>
> Milton Friedman (1953), *Essays in Positive Economics*

Moral philosophers have traditionally seen the field of ethics as practical in nature, as its goal is to provide us with the knowledge of what is right and what is wrong to do, which ought to be binding in nature. As an example of a view from philosophy, see David Hume's claim:

> "*Philosophy is commonly divided into speculative and practical; and as morality is always comprehended under the latter division, 'tis supposed to influence our passions and actions*"
>
> David Hume (1739), *A Treatise of Human Nature*

Hume's conception of human action is closer to Marshall's, and contrasts with that of Friedman's, who, in principle, saw positive economics as independent of normative judgments. Milton Friedman has been one of the greatest contributors to the methodology of the economic science, but one may wonder: *what if* David Hume was right? *What if*, nonetheless the well-founded methodological concerns that we may have, people's normative judgments about the strategies available to them in the decision situation they face influence their actual play in the game? If this is the case, then an economic model that aims to describe the causes of actions will be incomplete without considering such normative judgments. This thesis lies at the core of this debate, and its objective is to empirically explore whether normative judgments do, indeed, influence people's behaviour, and hence if they shall form part of what we know as positive economics, providing two such models – that of blame avoidance and praise seeking – as defined above.

Neoclassical economists, as a first step into generating a model of human action to be applied in economic settings, envisaged a narrow conception of human motivations for action, being the material payoff resulting from a strategy combination the primary determinant of people's utility. Hence, the well-known predictions of play in social settings: economists predicted people to offer nothing in ultimatum or dictator games, to defect in social dilemmas, and more important to the theme of the thesis, to under-supply public goods in the absence of self-regarding incentives to do so. Any evidence against those predictions was seen as an anomaly, and, indeed, Economists in the 1980's documented a wide range of such anomalies. As a response to them behavioural economics broadened the preference domain of a subject's utility, enlarging one's own motivational space to include inequality aversion, reciprocity, impure altruism, social efficiency, and maximin concerns, among others, thus circumventing the inconsistencies of the narrowly conceived vision that neoclassical economics had of the human nature.

Nowadays, it is widely accepted that people have such other-regarding preferences, but one may wonder whether ethical judgments, characteristic of their binding nature and their obligatory mandates, could also be a potential explanation for the altruistic behaviour displayed by people. Both views do not need to be orthogonal to each other: one can credibly argue that normative judgments induce a person to develop strong aversions against violating the ethical code derived from such moral judgments. However, what I want to contend is that both have different implications for the motivations underlying a subject's choice: while the broadened preference domain approach is self-centred and considers actions as flowing from how certain social properties influence a person's own utility, an approach based on moral judgments captures a disinterested approach to decision making. Namely, that a person acts not because inequality, or other social characteristics, pains them, but because they judge inequality as morally wrong; possibly, but not only, because it generates a pain in others, regardless of how much it generates in them. This thesis, by examining the extent to which moral judgments influence behaviour through the theoretical constructs of blame avoidance and praise seeking mentioned earlier, sheds new light on the implications of the altruistic nature of humans.

To achieve all the aims discussed earlier, this thesis contains seven chapters. The first chapter is the one you are reading, and presents an introduction to the aims, objectives and work done. The second chapter presents the first experiments that were carried out in the thesis. In short, I investigated the moral evaluations of impartial spectators regarding play in a three-person, simultaneous move, one-shot public goods game. Impartial spectators rated one of the three players in several scenarios, on a scale from -50 (extremely bad) to +50 (extremely good), were I manipulated (i) the framing of the decision situation as either a give-some or a take-some public good; (ii) the effective contribution of the judged player towards the public good; (iii) the effective average contributions of the non-judged group members towards the public good; and (iv) the dispersion in the effective contributions of the non-judged group members towards the public good. I find that factors (i) to (iv) influence the moral judgment our subjects ascribe to the judged person: (i) free riding (resp. half contribution) is seen as more blameworthy (resp. more praiseworthy) in give-some than in take-some public goods; (ii) free riders are judged as blameworthy, and contributors are judged as praiseworthy; (iii) moral judgments ascribed to free riders and half-contributors are decreasing in the average contribution of the non-judged group members; and (iv) free riding (resp. half contribution) is seen as less blameworthy (resp. less praiseworthy) when the contributions of the non-judged group members are different than when they coincide.

Chapter three studied, within a two-person, simultaneous move, one-shot public goods game, whether *blame avoidance* and *praise seeking* were determinants of (i) cooperation attitudes, as elicited by the strategy method presented in Fischbacher et al (2001), when controlling for the contribution of the other group member; and (ii) unconditional contributions, when controlling for the ABC method as advanced in Fischbacher and Gächter (2010). I studied the empirical validity of both theories for (i) and (ii) for Provision and Maintenance public good problems (or give-some and take-some versions of public goods problems). I find that (i) blame avoidance, but not praise seeking, influence cooperation attitudes of Provision and Maintenance problems even when controlling for the contribution of the other person; and (ii) that blame avoidance and praise seeking influence unconditional contributions to both Provision and Maintenance public goods. More specifically, I find that blame avoidance and praise seeking have both direct and indirect effects (i.e., through influencing cooperation attitudes, which influence the ABC method's prediction) on unconditional contributions. Both effects are more pronounced in Maintenance than in Provision public goods problems.

Chapter four embeds blame avoidance and praise seeking into a more general framework, the MRC framework, which allows me to present the theories more formally. I use the newly developed framework to test the two theories against six canonical models of economic decision-making: material selfishness, inequality aversion, reciprocity, social efficiency, maximin, and spite. The test is performed regarding contribution attitudes, as defined above, towards a social dilemma and a common interest game version of a public goods problem. A *common interest game* is defined as a public good game where the marginal per capita return to contributions is greater than one, thereby making full contribution both the selfish and the social optimum. We find that blame avoidance and praise seeking add to our understanding of cooperation attitudes in both social dilemmas and common interest games, as they predict data beyond what can be predicted by the other canonical models. Blame avoidance deserves special mention, as it is within the set of the best performing theories predicting cooperation attitudes of social dilemmas and common interest games, accompanied by models incorporating inequality aversion and maximin motives.

Chapter five concludes the core of the thesis, summarising the main results of the three main chapters (two, three, and four) of the thesis. Chapter six provides the reader with three appendices, one for each of the core chapters of the thesis. The reader will be referenced in the main text to the relevant parts of the appendices. Finally, chapter seven provides the references for all the works cited in the thesis.

# Chapter 2. The moral perception of players in social dilemmas

## 2.1. Introduction

In this chapter we report the findings of two studies that aim to investigate the moral significance of free riding and cooperation in social dilemmas from an impartial perspective. Furthermore, we study whether two newly developed theories of morality within the moral psychology literature, "*Moral Foundations Theory*" (e.g., Haidt and Joseph, 2004) and "*Morality As Cooperation*" (e.g., Curry, 2016), can explain individual differences in people's understanding of the moral significance of free riding and cooperation from that impartial perspective[1].

Cooperation is conjectured to be important from a moral perspective, as judged by several moral theories that argue for a link between both concepts (see Curry, 2016 for a review of such link along with Haidt and Kesebir's, 2010 and Curry et al's., 2019 definitions of morality as a regulator of social behaviour). Yet, and as argued by some moral psychologists, a vast amount of research output has put its focus on studying the moral perceptions of scenarios that abstract from everyday life situations such as cooperation issues (see, for instance, Bauman et al, 2014, and Graham, 2014 for critiques along these lines). Adding to such external validity concerns of studies within moral psychology, some authors state that trolley problems, commonly used to pit utilitarian against deontological moral judgments, are not satisfactory to capture an essential feature of utilitarian philosophy, viz., its impartial beneficence; asking for new experimental paradigms that can adequately test that defining feature of utilitarian

[1] A detailed coverage, or review, of Moral Foundations Theory and Morality As Cooperation theory lies outside the scope of the chapter and the thesis. Rather, we are strictly interested in whether those theories can explain our subjects' moral judgments of free riding and cooperation in social dilemmas. Hence, in the main text of the chapter we focus on testing the theories, and we relegate a presentation of a summary of both theories to Appendix A.3. Additionally, the spelling of the latter theory is normally *Morality-as-Cooperation,* rather than our usage *Morality As Cooperation*. However, as we use the acronym MAC later to refer to the theory we considered capitalising the second word as a natural step.

philosophy (see, most notably, Kahane, 2015; Kahane et al, 2015; Kahane et al, 2018; and Everett and Kahane, 2020).

Those criticisms form an important part of our motivation to study the moral judgments of social dilemmas for two main reasons. First, social dilemmas are, perhaps, the most canonical form of cooperation present in daily life and have arguably been the most used games to investigate cooperation in several disciplines. Hence, they do not lack external validity – they are, in an important sense, the most common decision situations of cooperation faced by humans throughout history. Second, and unlike trolley problems, cooperation in social dilemmas has the potential to capture utilitarianism's impartial beneficence. To see this, note that, in trolley problems, the utilitarian solution entails sacrificing others' wealth to achieve a higher societal wealth (e.g., killing one person to save several other persons). This solution captures the darker side of utilitarianism: the side that prescribes harming others when such harm is smaller than the aggregate benefit to society. Social dilemmas capture a fundamentally different feature of utilitarianism, as the utilitarian solution (i.e., full cooperation) prescribes sacrificing one's own wealth to increase the total wealth in a society. That is, the utilitarian solution to social dilemmas captures benefiting others, even at one's own cost, to promote the greatest overall wealth in a society. Both solutions strike us as radically different, and the increasing interest within moral psychology to investigate the positive side of utilitarianism makes social dilemmas a fruitful tool for such an endeavour. Although in this chapter we do not explicitly compare utilitarianism with deontology, or other moral theories, as an explanation for our moral judgments, we considered the development of an scenarios-based elicitation tool measuring moral judgments of social dilemmas to be important to the discipline in their search for different methods to study features of utilitarianism that have, hitherto, received little attention.

In this chapter we use a tool developed by Cubitt et al. (2001) to elicit the impartial moral views regarding behaviour in social dilemmas. In both studies we present subjects some scenarios based on a three-person, simultaneous game, one-shot version of a social dilemma. Each scenario shows a description of the decision situation and the actions of each of the people involved in the decision situation (henceforth, group members). We ask our subjects to judge one of the group members from an impartial spectator viewpoint (that is, taking the position of a person not taking part in the decision situation) on a scale from -50 (extremely bad) to +50 (extremely good). We manipulate each of four key variables between scenarios orthogonally: (i) the total contribution to the public good of the judged group member (as a consequence of their action); (ii) the average contribution to the public good of the non-judged group members; (iii) the dispersion in the contribution of the non-judged group members; and (iv) the frame of the decision situation (Give frame for some scenarios; Take frame for other scenarios). This experimental design allows us to clearly identify the effect of each of the four variables in our subjects' impartial moral judgments of scenarios of social dilemmas.

Additionally, we present our subjects the Moral Foundations Questionnaire (Study1) and the Morality As Cooperation Questionnaire (Study 2), which aim to capture the strength of several concepts (henceforth, foundations) in a person's own conception of morality. We use the scores in all foundations to investigate whether individual differences in the importance of each foundation for our subjects' conception of morality explains, to some extent, the individual variation of our subject's impartial moral judgments, elicited as described in the previous paragraph.

We find in Study 1 – and replicate in Study 2 – that selfish and cooperative behaviour in social dilemmas trigger strong and vastly different impartial moral judgments: free riding triggers condemnation and cooperation triggers moral praise. More specifically, the four variables that we manipulated between scenarios are key determinants of the impartial moral judgments of our scenarios, and their specific effect can be summarised as follows. First, impartial moral judgments are influenced by the contribution of the judged group member, evidenced by the fact that the judged person is praised more the higher their contribution. Second, impartial moral judgments are influenced by the average contribution of the non-judged group member, as the moral ratings decline whenever the contribution of the judged person is lower than the average contribution of the non-judged group members. Third, impartial moral judgments are influenced by the dispersion in the contribution of the non-judged group members, as free riders are perceived as less blameworthy, and half contributors are perceived as less praiseworthy, when the non-judged group members contribute differently. And fourth, impartial moral judgments are influenced by the framing of the decision situation, as free riders are perceived as more blameworthy, and half contributors are perceived as more praiseworthy, in give-some social dilemmas than in take-some social dilemmas. The foundations of Moral Foundations Theory and Morality As Cooperation theory can capture individual differences in moral judgments, although their relative importance is smaller than the variables manipulated across scenarios.

This chapter makes two important contributions to economics and moral psychology. First, we build on Cubitt et al (2011) in that we study impartial moral judgments of social dilemmas. However, we extend their work in such a way that allows us to study not only the moral dimension of free riding but also the moral dimension of half and full contribution. The works within economics that specifically study the relation between morality and cooperation are scarce (but see, most notably, Dal Bó and Dal Bó, 2014; Hauge, 2015; and Cappelen et al, 2019), and we contribute to that literature by providing an overall picture of how a wide range of actions within a canonical cooperation problem are perceived by impartial spectators. Second, we contribute to the literature in moral psychology in that we bring Moral Foundations Theory and Morality As Cooperation theory directly to the test in their most natural setting. That is, our study allows us to explore whether theories that were generated to capture the relation between cooperation and moral judgments are, indeed, able to explain people's moral judgments towards violations of (i.e., free riding) or engagement in (i.e., contribution) cooperation.

The chapter proceeds as follows. Section 2.2 presents the tool we develop to elicit the impartial moral judgments of several scenarios to subjects. Section 2.3 presents the experimental design, and results, of Study 1. Section 2.4 presents the experimental design, and results, of Study 2. Section 2.5 concludes by summarising the main results and outlining future potential avenues of research in lieu of our results.

## 2.2. Moral Evaluation of Social Dilemmas (MESD)

Cubitt et al (henceforth CDGK, 2011) introduced an experimental survey method to elicit subjects' moral perceptions of free riders in social dilemmas. More specifically, they presented subjects several scenarios regarding a public goods game involving two people, and ask them to judge, from an impartial spectator viewpoint, the morality of one of the two group members. The Moral Evaluation of Social Dilemmas (henceforth, MESD) we use is based on the elicitation procedure developed in CDGK (2011), as we retain the roles of the experimental subject as an impartial spectator, and the judged and non-judged group members as the basis of the scenarios.

We extend (and modify) their design in two important ways. First, we vary the actions of the judged person across the scenarios presented to subjects. This allows us to study how the moral perception ascribed to a group member changes with their action in a public goods game. And second, scenarios involve two, rather than one, non-judged group members. This allows us to study whether the heterogeneity in the non-judged group members' actions influences the moral perception ascribed to the judged players. Before introducing the scenarios, we describe the decision situation on which the scenarios are based.

### 2.2.1. *The decision situation*

The scenarios are based on the following version of a three-person, simultaneous, one-shot public goods game. Each group member controls 20 tokens. Each group member must allocate their tokens between a group project and their private account, and the allocation decision is restricted to be an element of $\{0,10,20\}$. Decisions are made simultaneously and in private.

Each group member receives $1 for each token they allocate to their own private account; and each token allocated to the group project yields $0.5 for each group member. Hence, each token allocated to the project generates more than $1 aggregate wealth in the group (as $1.5 > $1), but each group member gains more payoff by allocating each token to their private account, as $0.5 < $1. The monetary consequences derived from the allocation decisions are common knowledge and the same for every group member.

We considered two different frames of the social dilemma: Give and Take. The only difference between frames is the initial allocation of players' tokens. In the Give frame, each group member initially has 20 tokens in their private account whereas the group project has 0 tokens; and each group member has to decide how many tokens to *contribute* to the group project. In the Take frame, the group project initially has 60 tokens whereas each group member has 0 tokens in their private account; and each group member's decision is to decide how many, up to 20 tokens, to *withdraw* from the project.

For the remainder of the chapter, we refer to the number of tokens contributed to (in the Give frame) or left in the group project (in the Take frame) as that person's *effective contribution*. Using the concept of effective contribution allows us to isolate the effect of the framing of the game from the monetary consequences in the moral perceptions of the different scenarios we study.

### 2.2.2. *The scenarios*

Each scenario presents the description of the decision situation and an '*ending*' to the decision situation. The ending of the decision situation consists of the allocation decisions of the three group members. We label the judged group member as Person A and the non-judged group members as Persons B and C. Our subjects' task was to rate the morality of Person A between -50 (extremely bad) to +50 (extremely good) in each scenario we presented to them from an impartial spectator stance. That is, our subjects did not play the game and, hence, had no stakes in the decision situation.

We first introduce some notation to be able to define the scenarios in a compact way. Let $f \in \{Give, Take\}$ be the frame of the decision situation, $C_A, C_B, C_C \in \{0,10,20\}$ be the effective contribution levels of the three group members, and $C_{BC} \in \{0,10,20\}$ be the average effective contribution of Persons B and C. Additionally, let $d \in \{Equal, Unequal\}$ take the value $Equal$ when Person B and Person C's effective contributions coincide ($C_B = C_C$) and $Unequal$ otherwise (i.e., when $C_B \neq C_C$). A scenario is, then, defined by a quadruple of the form $\langle f, C_A, \bar{C}_{BC}, d\rangle$. Since each group member's effective contribution is an element of $\{0,10,20\}$, it follows that both $C_{BC} = 0$ and $C_{BC} = 20$ require $d = Equal$, whereas $C_{BC} = 10$ is also compatible with $d = Unequal$. As we do not give any information to distinguish between Person B and Person C, except possibly their effective contributions, we focus on combinations, rather than permutations, of $C_B$ and $C_C$.[2] The scenarios we study, in terms of a generic frame *f*, are shown in Table 2.1:

The MESD consists of 24 scenarios, 12 per frame. For the remainder of the paper, when we refer to the scenarios we will only include the term $d$ when allocations of Person B and Person C are unequal (i.e., $d = Unequal$). Thus, for instance, with

[2] Note that, for instance, the scenario $\langle Give, 0{,}10, Unequal\rangle$ can be achieved with either $\langle C_B = 20, C_C = 0\rangle$ or $\langle C_B = 0, C_C = 20\rangle$. To avoid redundancy, in our studies we only present scenarios involving one of the two possibilities to the subjects. The specific permutation for each combination of $C_B$ and $C_C$ chosen for constructing our scenarios is kept constant across treatments.

$\langle f, 10{,}10\rangle$ we refer to the scenario where $C_B = C_C = 10$ and with $\langle f, 10{,}10, Unequal\rangle$ we refer to the scenario where $C_B = 20 > C_C = 0$.

**Table 2.1.** *Scenarios included in the task provided to subjects facing frame f*

| | | $C_{BC}$ | | |
|---|---|---|---|---|
| | | 0 | 10 | 20 |
| $C_A$ | 0 | $\langle f, 0{,}0, Equal\rangle$ | $\langle f, 0{,}10, Equal\rangle$<br>$\langle f, 0{,}10, Unequal\rangle$ | $\langle f, 0{,}20, Equal\rangle$ |
| | 10 | $\langle f, 10{,}0, Equal\rangle$ | $\langle f, 10{,}10, Equal\rangle$<br>$\langle f, 10{,}10, Unequal\rangle$ | $\langle f, 10{,}20, Equal\rangle$ |
| | 20 | $\langle f, 20{,}0, Equal\rangle$ | $\langle f, 20{,}10, Equal\rangle$<br>$\langle f, 20{,}10, Unequal\rangle$ | $\langle f, 20{,}20, Equal\rangle$ |

## 2.3. Study 1

### 2.3.1. *Methods*

#### 2.3.1.1. *Procedures*

We presented two different tasks to our participants. One of the tasks was a set of 12 scenarios from the MESD we developed. We presented to each subject all the scenarios involving a given frame. That is, we manipulated $f$ between subjects and $C_A$, $C_{BC}$ and $d$ within subjects. The other tasks we presented to our subjects was the Moral Foundations Questionnaire developed by Graham et al (2011), to which we added the additional material developed in Iyer et al (2012). The Moral Foundations Questionnaire was built to operationalize Moral Foundations Theory (henceforth, MFT), by Haidt and Joseph (2004) and subsequently refined in Haidt and Joseph (2007), Haidt and Graham (2007), Haidt and Graham (2009), Graham et al (2009), and Graham and Haidt (2010). Put shortly, MFT presents a theoretical framework that depicts moral judgments as (i) mainly coming from intuition; (ii) culturally variable; (iii) a path towards social regulation; (iv) fed by plurality of aspects; and (v) serving the purpose of binding people in communities (see Haidt, 2007 and 2013 for reviews, and 2012 for a book-length accessible summary of the theory).

To capture the plurality of aspects considered in people's morality, MFT proposes the concepts of Harm/Care, Fairness/Reciprocity, Ingroup/Loyalty, Authority/Respect, and Purity/Sanctity[3]. Additionally, Iyer et al (2012) propose Economic and Lifestyle Liberty to be further aspects important to moral judgments. For compactness, in the remainder of the chapter we refer to those concepts as the *moral foundations* of Harm, Fairness, Loyalty, Authority, Purity, Economic, and Lifestyle Liberty[4]. Crucially, Graham et al's (2011) Moral Foundations Questionnaire elicits a score per each of these moral foundations, ranging from 0 to 5. The score is to be interpreted as the importance of a given aspect for one's own conception of morality, a higher score denoting a higher relevance of a given aspect to one's own conception of morality. We use the scores per each foundation at the individual level to test whether, and, if so, how, moral foundations proposed by MFT influence the moral judgments of social dilemmas, as elicited by MESD.

We manipulated the order in which we presented the two tasks between subjects. This resulted in a 2 × 2 between-subjects design, where the framing of the decision situation and the order of the two tasks were the treatment variables. Subjects were randomly allocated to one of the four treatments and were allowed to participate only once.

After presenting a description of the decision situation, and before subjects could make the moral ratings of the 12 scenarios, they had to answer several control questions to assess their understanding of the determination of monetary consequences in the scenarios[5]. Only subjects that answered them correctly were able to continue with the experiment. This allowed us to make sure that subjects finishing the experiment understood the monetary and distributional consequences of different

[3] Here, and in what follows, we fix the notation of the moral foundations to be the one presented in Haidt and Joseph (2007).

[4] The concept of a foundation is a difficult one, as MFT feeds from several disciplines, such as evolutionary psychology and cultural psychology. We will not define the concept in detail here, but, put shortly, a module is a region within the mind that contains innate information, is fast, and which arose favoured by evolution (for a more detailed exposition, see Haidt and Joseph, 2007, section 3.4). Hence, moral foundations can be seen, more coarsely, as an innate predisposition to consider aspects as innately relevant for our moral judgments.

[5] The frame of the description of the decision situation, of the control questions and of the scenarios task was the same within subjects.

combinations of effective contributions before they made their judgments. Hence, and beyond any mistakes in judgments that could have arisen from confusion, this design feature ensured that insensitivity of our subjects' impartial moral judgments with respect to $C_A$, $C_{BC}$ and/or $d$ was intentional.

We paid subjects a flat fee of $3 after they completed the study. Payment was not conditional on the responses of the subjects, to avoid confounding their moral perceptions of the scenarios with beliefs about which responses would be better rewarded (for a discussion of the rationale of this incentive scheme, see CDGK, 2011, p. 257). The instructions of the experiment are provided in Appendix A.1 of the thesis.

#### 2.3.1.2. *Participants*

We recruited 398 participants on Amazon Mechanical Turk (average age: 35.45 years; 44% females; 53% liberals; and 38% religious)[6]. Subjects of Study 1 replicate the main empirical regularities of MFT: differences in scores between liberals and conservatives, similar factor loadings of exploratory factor analysis and similar qualitative results from confirmatory factor analysis – See Appendix A.2 of the thesis.

### 2.3.2. *Results*

#### 2.3.2.1. *Moral Judgments*

We use the *Moral Evaluation Functions* (henceforth, MEF) as the main way to analyse our data. CDGK (2011) defined the MEF as the average moral judgment ascribed to free riders expressed as a function of the actions of the non-judged group member. However, our study differs from theirs in two important ways. First, our scenarios study the judged person (Person A) in situations where Person A does not necessarily free ride. Second, our scenarios are based on a group with two non-judged group members (Persons B and C) rather than one. As a natural step, we revise CDGK's (2011) definition to make it more general. We redefine the MEF of $C_A$ as the

[6] We asked each subject to fill a sociodemographic questionnaire within the experiment. Data reported here refers to subjects' self-reported statements. For more detail on all the questions within the sociodemographic questionnaire, and the specific scale of each of the questions, one can refer to the experimental instructions provided in appendix A of the thesis.

mean moral rating ascribed to Person A as a function of the effective average contributions of the non-judged group members.

We report in Figure 2.1. the Moral Evaluation Functions of free riders ($C_A = 0$, left panel), half contributors ($C_A = 10$, centre panel) and full contributors ($C_A = 20$, right panel) for the Give and Take treatments independently (Give treatments: solid lines; Take treatments: dashed lines), plotting each average moral judgment with its corresponding 95% confidence interval. Additionally, we plot with the symbol "X" the average moral rating of scenarios where Person B and C's effective contributions differ ($d = Unequal$. Give treatments: black cross; Take treatments: grey cross). The horizontal axis gives the average effective contribution of the non-judged group members ($C_{BC}$) and the vertical axis our subjects' impartial average moral rating ascribed to Person A. For reference, in each panel we draw a dashed line at a moral rating of 0, representing a benchmark of moral neutrality where neither scenario is of moral significance.

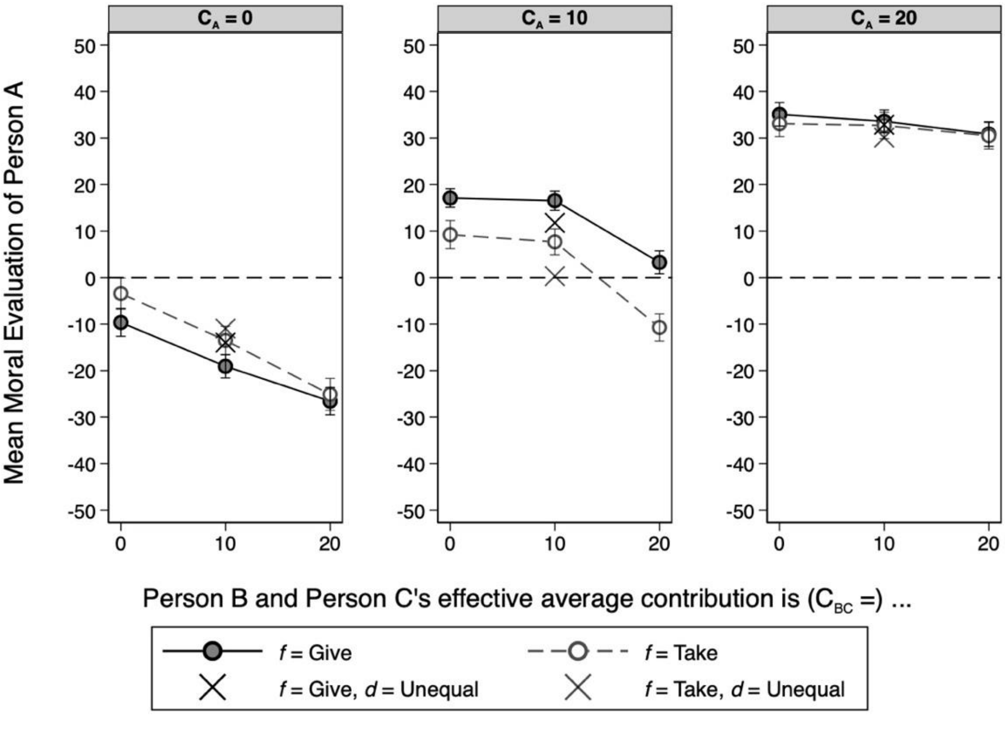


**Figure 2.1.** *Mean Moral Evaluations (of impartial spectators) ascribed to Person A - Study 1*

We observe five behavioural patterns in Figure 2.1. First, and most importantly, social dilemmas have a moral dimension in the eyes of impartial spectators. In each of the three panels, the MEF's differ markedly from the moral neutrality benchmark (the dashed lines at 0), providing support for the claim that our subjects view scenarios of social dilemmas as of moral significance in their impartial moral judgments of Person A.

Second, our subjects' impartial moral judgments ascribed to Person A are, on average, increasing in $C_A$ (i.e., Person A's own contribution): effective free riders are judged as blameworthy, and effective contributors are judged as praiseworthy; and subjects judge Person A as more praiseworthy the higher their effective contribution. These patterns emerge regardless of the frame and regardless of what Person B and Person C's effective contributions are.

Third, the impartial moral judgments ascribed to Person A crucially depend on how $C_A$ compares with $C_{BC}$ (i.e., the average effective contribution of Persons B and C). Whenever $C_A \geq C_{BC}$, the slope of the MEF's is flat. In contrast, the MEFs are negatively sloped when $C_A \leq C_{BC}$.

Fourth, the moral perceptions of social dilemmas are influenced by $d$ (i.e., the heterogeneity in the effective contributions of the non-judged group members). In scenarios where $C_{BC} = 10$, we can observe that free riders are judged to be less blameworthy, and half contributors are judged to be less praiseworthy in the scenarios where $d = Unequal$ than in scenarios where $d = Equal$. The lower condemnation of free riders is especially prominent in the Give frame, whereas the lower praiseworthiness of half-contributors is as equally marked in both frames.

Fifth, the moral perception of social dilemmas is influenced by $f$ (i.e., the framing of the decision situation). More specifically, the influence of framing is different at each effective contribution level of Person A: free riders are judged to be more blameworthy in the Give than in the Take frame, half contributors are judged to be more blameworthy in the Take than in the Give frame, and full contributors are judged to be equally praiseworthy in both frames.

We investigate not only the average impartial moral judgments but also the level of dispersion of our subjects' moral views of social dilemmas, that is, the degree of individual differences on people's impartial moral ratings of Person A. To report the heterogeneity of our subjects' impartial moral judgments of the scenarios in the Give and Take treatments, Figure 2.2. shows the violin plots of all the scenarios for which $d = Equal$. The top three panels contain the violin plots of the Give scenarios, and the bottom three panels contain the violin plots of the Take scenarios. Both the top and bottom rows follow closely the structure of Figure 1.1. (i.e., each row contains three panels, each panel contains all scenarios for a given effective contribution level of Person A, and panels are ordered, from left to right, in increasing order of $C_A$).

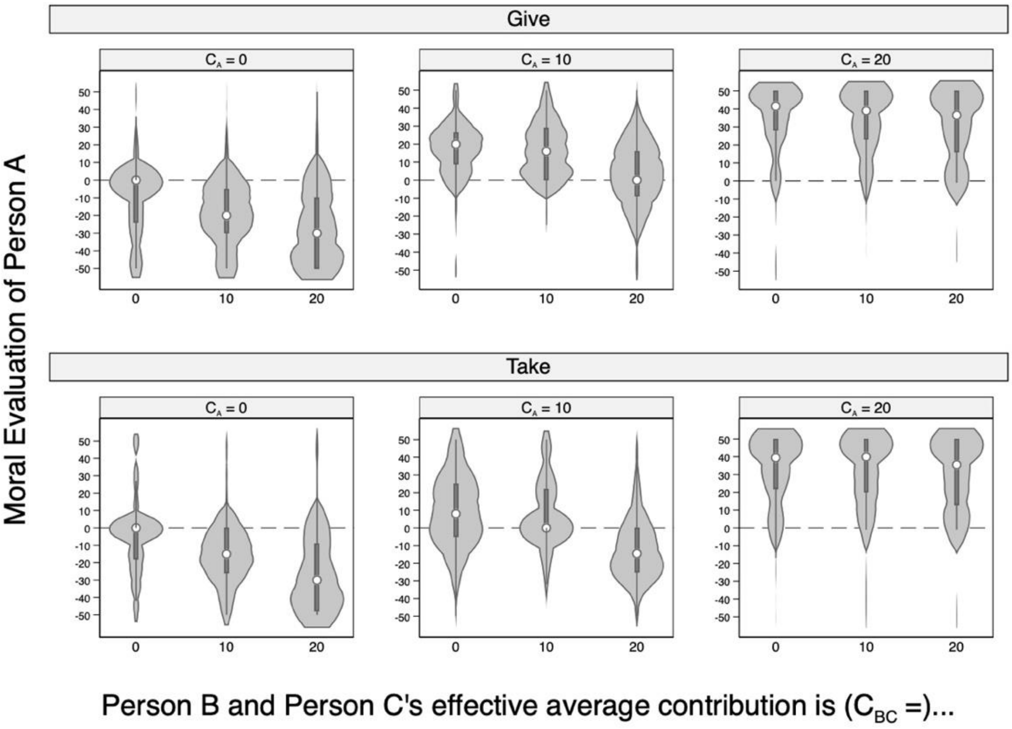


**Figure 2.2.** *Violin plots of the moral judgments of scenarios - Study 1*

Each violin plot shows the median moral judgment (white dot), the interquartile range (dark bar), the whiskers (dark line) and a measure of density (grey area) of the moral judgments per scenario (see Hintze and Nelson, 1998, for a detailed description of violin plots). We observe that in most cases the density of moral judgments spread over a substantial part of the range of moral ratings for both Give (top row) and Take (bottom row) scenarios. This shows that, at the individual level, people do not converge to a singular moral rating but, rather, that they perceive scenarios differently.

Finally, and given the existence of a substantial degree of dispersion in the impartial moral judgments of scenarios as captured by Figure 2.2., we wanted to explore the moral judgments from a more qualitative perspective in the search of a plausible framework that can capture such individual differences. For that, we calculate the proportion of different shapes of MEF's for different levels of $C_A$. This allows us to observe if there is a systematic structure underlying the dispersion of moral judgments; and, if so, whether that structure depends on the effective contribution level of the judged Person (i.e., $C_A$).

We, first, provide below precise definitions of the five different shapes of MEF's that we consider: those being *flat*, *decreasing*, *increasing*, *triangle*, and *reverse triangle* shapes of MEF's. For compactness of the definitions, we fix $m_i$ to be the impartial moral judgment of subject $i$:

- A MEF of $C_A$ is categorised as *flat* for a given $f$ iff $m_i(\langle C_A, 0, f\rangle) = m_i(\langle C_A, 10, f\rangle) = m_i(\langle C_A, 20, f\rangle)$
- A MEF of $C_A$ is categorised as *decreasing* for a given $f$ iff $m_i(\langle C_A, 0, f\rangle) \geq m_i(\langle C_A, 10, f\rangle) \geq m_i(\langle C_A, 20, f\rangle)$, with at least one inequality holding strictly.
- A MEF of $C_A$ is categorised as *increasing* for a given $f$ iff $m_i(\langle C_A, 0, f\rangle) \leq m_i(\langle C_A, 10, f\rangle) \leq m_i(\langle C_A, 20, f\rangle)$, with at least one inequality holding strictly.
- A MEF of $C_A$ is categorised as *triangle* for a given $f$ iff $m_i(\langle C_A, 0, f\rangle) < m_i(\langle C_A, 10, f\rangle) > m_i(\langle C_A, 20, f\rangle)$.
- A MEF of $C_A$ is categorised as *reverse triangle* for a given $f$ iff $m_i(\langle C_A, 0, f\rangle) > m_i(\langle C_A, 10, f\rangle) < m_i(\langle C_A, 20, f\rangle)$.

Table 2.2. reports the proportion of each shape of MEF for all effective contribution levels of Person A separately for Give and Take treatments. The first two columns (resp. the central two columns; resp. the last two columns) present the proportion of each shape of MEF for $C_A = 0$ (resp. $C_A = 10$; $C_A = 20$). The last row reports the chi 2 tests to assess whether the distribution of the adherence rate of the shape of MEF's differ across frames, but we do not find any evidence to support that claim. Hence, the dispersion of moral judgments of scenarios, as least when looked from the *shapes of the MEF's* perspective, does not vary across frames.

**Table 2.2.** *Proportions of the different shapes of the MEF's of each level of $C_A$ for each frame - Study 1*

| | $C_A = 0$ | | $C_A = 10$ | | $C_A = 20$ | |
|---|---|---|---|---|---|---|
| | Give | Take | Give | Take | Give | Take |
| Flat | 22.12% | 22.11% | 13.46% | 13.68% | 36.06% | 37.37% |
| Decreasing | 58.65% | 56.84% | 37.98% | 47.89% | 27.40% | 22.63% |
| Increasing | 1.92% | 4.74% | 9.62% | 5.79% | 14.42% | 13.68% |
| Triangle | 9.13% | 12.63% | 31.73% | 26.32% | 10.10% | 11.05% |
| Rev. Triangle | 8.17% | 3.68% | 7.21% | 6.32% | 12.02% | 15.26% |
| Chi 2 test | $\chi^2_{(4)} = 6.91; p = 0.14$ | | $\chi^2_{(4)} = 5.27; p = 0.26$ | | $\chi^2_{(4)} = 1.84; p = 0.76$ | |

Table 2.2 shows that there is a substantial adherence to different shapes of MEF at each level of $C_A$. And, more importantly, we notice that the adherence to a given shape of MEF varies with $C_A$. For instance, we observe the decreasing MEF has its greatest adherence in free riding; and the adherence to it drops substantially the higher $C_A$. In contrast, the increasing MEF has the greatest adherence at the highest level of $C_A$, and its adherence decreases the lower the level of $C_A$. A triangle MEF shows its greatest adherence at $C_A = 10$, and a flat MEF is most common at $C_A = 20$, and at $C_A = 0$. These patterns emerge in both Give and Take frames of social dilemmas.

Taken together, these findings suggest that the dispersion in moral judgments can be captured by the different frequencies of shapes of MEF's. But, more importantly, the different shapes of MEF's at different levels of $C_A$ suggest that the dispersion of the data is dependent on $C_A$. This is perhaps better exemplified by referring to the distribution of joint MEF's[7]: less than one-third of our subjects (around 31% in either frame) have the same shape of MEF at all contribution levels, and most that do, hold either a decreasing (around 17% out of the total subjects in either frame) or a flat (around 11% of subjects in either frame) MEF at all levels of $C_A$. Between one-fifth and one-quarter of subjects (around 22% in either frame) switch from a decreasing MEF of free riders to a non-decreasing MEF of half contributors, being the most common pattern within that switch the one involving a decreasing MEF of free riders

[7] As we have five different shapes of MEF's and three different levels of $C_A$, this implies the potential of 125 potential different joint MEF's. Whilst we do not report the full distribution of the 125 joint MEF's as most have little to no adherence, in the main text we discuss the most common forms of joint MEF's that our subjects endorsed.

and a triangle shape of half contributors (around 15% subjects of either frame hold this pattern).

### 2.3.2.2. *Moral Foundations Theory*

To analyse the effect of moral foundations on the moral judgments of the MESD scenarios we run the following OLS regression seven times, one per each of the moral foundations captured by MFT, clustering in each the standard errors at the individual level to control for the dependency of moral judgments within individuals:

$$(2.1) \quad m_{i,j} = \beta_1 + \sum_{j=2}^{24} \beta_j * D_j + \beta_{25} * z\_MF_i + \sum_{j=2}^{24} \beta_{24+j} * D_j * z\_MF_i + u_{i,j}$$

In equation (2.1) the dependent variable $m_{i,j}$ represents the moral judgment of subject $i$ in scenario $j$ from an impartial spectator stance; $u_{i,j}$ represents the error term in the regression equation; and the independent variable $z\_MF_i$ represents subject $i$'s score in the standardised transformation of a moral foundation score. That is,

$$(2.2) \qquad z\_MF_i = \frac{MF_i - \frac{\sum_{i=1}^{N} MF_i}{N}}{\sqrt{\frac{\sum_{i=1}^{N}\left(MF_i - \frac{\sum_{i=1}^{N} MF_i}{N}\right)^2}{N-1}}}$$

Hence, a value of -1 (resp. +1) in the $z\_MF_i$ variable implies that subject $i$'s score in the moral foundation $MF$ is exactly the value of one negative (resp. one positive) standard deviation from the sample mean in the moral foundation $MF$. For the remainder of the chapter, and for compactness and ease of understanding, we will refer to scores of -1 (resp. +1) as low (resp. high) scores in a moral foundation. Additionally, $D_j$ represents a generic dummy that takes the value 1 for scenario $j$ and 0 for all other scenarios. That is, there are 24 such dummies, one per each scenario outlined in the MESD section. We leave the scenario $j = \langle C_A = 0, C_{BC} = 0, f = Give \rangle$ as the baseline one and include the 23 remaining dummies in the regressions.

To examine how each of the moral foundations influences moral judgments, we calculate the predicted values of each of the 24 scenarios at $z\_MF_i = -1$ and $z\_MF_i = +1$, and plot those predicted moral judgments, with 95% confidence intervals, in different figures, one per each moral foundation[8]. This allows us to visually inspect whether, for each scenario, subjects with high and low scores in each foundation are predicted to have, on average, significantly different moral judgments in any of the scenarios we elicited via the MESD.

In the main text of the chapter, and to avoid unnecessarily inflating it with figures that look very similar, we only include the figures for the Harm and Fairness foundations, and we relegate the results of the other foundations to Appendix A.4.1. The decision to specifically keep Harm and Fairness within the main text is because they were, *ex ante*, the foundations more closely related to issues of cooperation. For instance, harming others can be interpreted, in a broader sense, as making another person worse-off with respect to the counterfactual of full contribution. Additionally, the Fairness foundation is commented by the authors of MFT (e.g., see Haidt and Joseph, 2007) to be the one more closely related to social dilemmas and free riding, so we considered it to be the most relevant to study within our Study.

To support our discussion of each foundation, both in here and the one ensuing in Appendix A.4.1, we provide the results of several tests of hypotheses regarding the regressions we run. We provide a rationale for and the specifics of the tests more formally in appendix A.5.

The three panels at the top of each of the subsequent figures (Figures 2.3 and 2.4) represent the predicted MEF's of Give scenarios for low and high scores in each foundation, while the three panels at the bottom represent the MEF's of the Take scenarios for low and high scores in each foundation. The panels are organised in a similar fashion as the ones presenting the MEF's earlier on. That is, each panel represents the average predicted moral judgments of all scenarios corresponding to a specific effective contribution level of the judged person (Person A); the vertical axis represents the predicted moral judgments from the regression, that can range between -50 (extremely bad) to +50 (extremely good); and the horizontal axis represents the effective average contribution of the non-judged group members (Person B and Person C). As before, the panels are displayed in increasing order of the effective contribution of Person A, being the panel corresponding to predicted moral judgments of scenarios

[8] It is those values (i.e., $z\_MF_i = -1$ and $z\_MF_i = +1$) the ones we use to refer to in the figures that follow as *low* and *high* scores of a given foundation, respectively. In order not to clutter the captions of each of the figures that follow with repetitive mathematical notation, we will use the terms *high* and *low* from now on to refer to 1 negative (resp. one positive) standard deviation from the standardised score in each foundation.

where $C_A = 0$ (resp. $C_A = 10$; resp. $C_A = 20$) the leftmost (resp. central; resp. rightmost) panel.

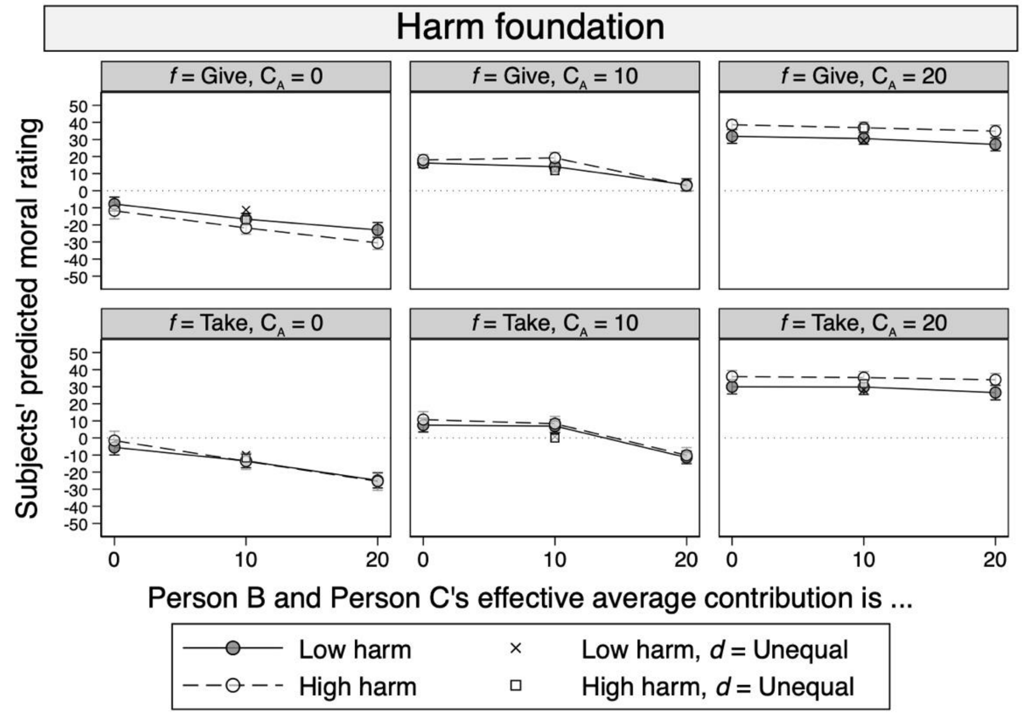


**Figure 2.3.** *Predicted MEF's evaluated at high and low scores of the Harm foundation - Top Panels: Give scenarios; Bottom Panels: Take scenarios*

We start our discussion with the predicted moral judgments of high and low scores of the *Harm foundation*, reported in Figure 2.3. When considering all 24 scenarios jointly, we find evidence of a small but significant effect of Harm on moral judgments ($F(24{,}397) = 1.60,\ p = 0.04$). More specifically, this effect is driven by the influence of Harm in the moral judgments of the Give scenarios (Give frame: $F(12{,}397) = 1.86$, $p = 0.04$; Take frame: $F(12{,}397) = 1.35$; $p = 0.19$). Overall, a higher score in the Harm foundation is associated with more blameworthy judgments of free riders in the Give frame (see top left panel) and more praiseworthy judgments of full contributors in both the Give and Take frames (see top and bottom right panels); albeit the effects are small. Additionally, we do not find evidence of Harm as a statistically significant moderator of framing effects ($F(12{,}397) = 0.90$, $p = 0.55$). Albeit not reaching statistical significance, there is a tendency for a high score (relative to a low score) in the Harm foundation to generate a more pronounced condemnation of free riders.

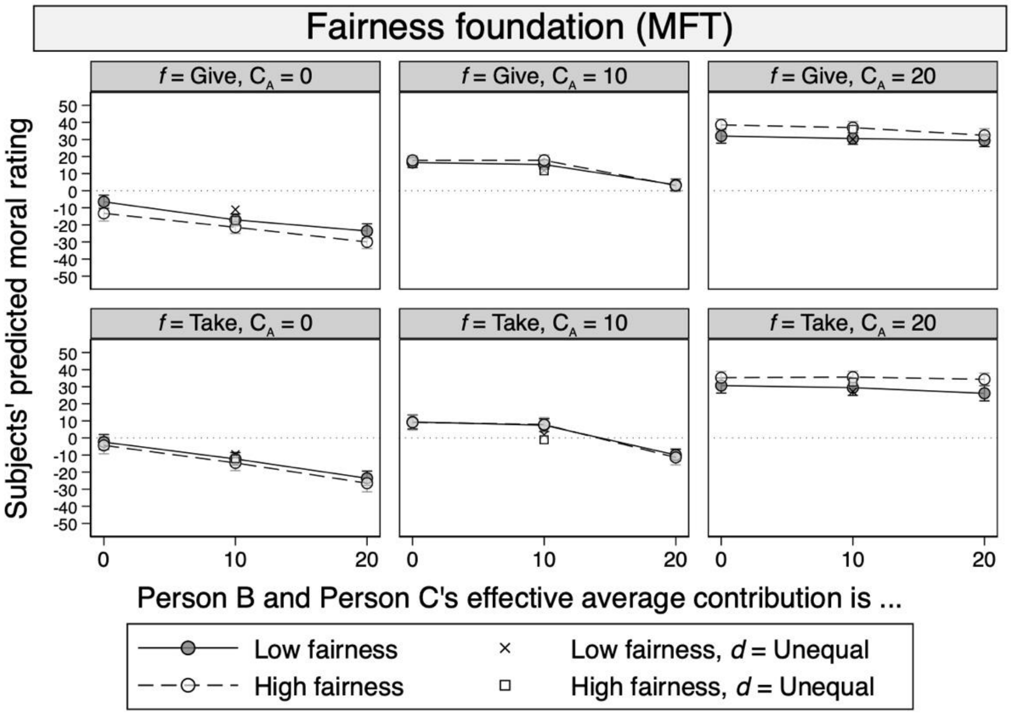


**Figure 2.4.** *Predicted MEF's evaluated at high and low scores of the Fairness (MFT) foundation - Top Panels: Give scenarios; Bottom Panels: Take scenarios*

Figure 2.4. reports the predicted moral judgments of all scenarios of high and low levels of the *Fairness foundation*. This is the foundation in MFT that is more closely related to social dilemmas, so we expected *ex ante* to be the one performing the best in driving moral judgments of MESD scenarios. However, we do not find evidence of a significant effect of Fairness on moral judgments; neither at the aggregate level ($F(24{,}397) = 1.34,\ p = 0.13$) nor when considering Give and Take scenarios independently (Give frame: $F(12{,}397) = 1.46, p = 0.14$; Take frame: $F(12{,}397) = 1.23$; $p = 0.26$). Even when the effects do not reach significance, they are in line with our expectation, viz., that a higher score in the Fairness foundation is associated with more blameworthy judgments of free riders in the Give frame (see top left panel) and more praiseworthy judgments of full contributors in the Give and Take frames (see top and bottom right panels). We additionally do not find evidence of Fairness being a moderator of framing effects, as it falls short of significance ($F(12{,}397) = 1.36$;

$p = 0.18$). Its qualitative effects, however, are similar to those of the Harm foundation.

To summarise briefly the results regarding the non-included foundations, we can make the following statements. First, we do not find any support for the Loyalty foundation as a significant driver of moral judgments of the Give or Take scenarios, neither support of the Loyalty foundation as a driver of the framing effects of the MEF's presented earlier. Second, the Authority and the Purity foundation have strong significant effects ($p < 0.05$) in predicting moral judgments of our scenarios, and only the purity foundation is a statistically significant driver of framing effects. However, their effect size in either case is small. Third, among the Liberty foundations the Economic liberty fares better, and plays an opposite role to that of the other foundations. Namely, that a higher score in the Economic liberty foundation decreases (rather than increases) the perceived blameworthiness of free riders in the Give frame.

### 2.3.3. *Discussion*

Overall, the data suggests that each action has its own moral space. Each action is perceived differently in terms of its morality: free riding is perceived as morally bad whereas contributions are perceived as praiseworthy. Moreover, we find that the circumstances that surround a scenario (what we label as $C_A$, $C_{BC}$, $f$ and $d$ in the scenarios) are crucial to determine the perceived morality of the judged group member from an impartial perspective. Additionally, we find large and substantial individual differences in moral judgments. This struck us as an interesting finding given that moral judgments are made from an impartial perspective, and thus provide evidence in favour of a subjectivist account of moral judgments: even when subjects do not have stakes in the social dilemmas and are given full information about the scenario to be rated, their impartial views of free riders, half contributors, and full contributors vary widely. Our results, thus, suggest that, on top of the level of relevant information being a driver of the dispersion of people's impartial moral judgments, as Konow (2009, pp.116-119) reports, individual differences are a crucial factor in impartial moral judgments of social dilemmas.

We also find that almost all moral foundations contribute to partially explaining the individual differences in moral judgments. More generically, we find that (i) a higher score in most foundations tends to increase the blame ascribed to free riders and the praise ascribed to full contributors in the Give frame; (ii) the effect of the foundations on the moral judgments of scenarios tends to be stronger in the Give frame; and (iii) that Economic liberty plays an opposite role to the other foundations, as a higher score in this foundation reduces the blame ascribed to free riders in scenarios of the Give frame. Additionally, none of the foundations appears to be a strong driver of the framing effects in the MEF's we report earlier. It is also worth noting that Harm and Fairness have the largest effects on moral judgments, although the statistical significance of the Authority and Purity foundations in the effects of moral judgments is stronger. This does not highlight that Authority and Purity are more important than Harm and Fairness in understanding individual differences of moral judgments. Rather, it implies that their estimates of an effect are more precise.

One conclusion we can take from our study is that the effect of the foundations is smaller than the effect of either variable manipulated within scenarios. This is not a surprise, as the information we provide to individuals is crucial to their understanding of behaviour and its implications (e.g., distribution of final outcomes). However, it highlights that there is scope for future research to identify different variables that aid to further advance our understanding of the individual differences in impartial moral judgments that we report in Figure 2.2 and Table 2.2, especially when it comes to our understanding of the sources of framing effects in moral judgments. It is partially with this in mind that Study 2 investigates how a newly developed theory in moral psychology, Morality As Cooperation theory, fares at explaining individual differences in moral judgments of social dilemmas.

## 2.4. Study 2

The purpose of this study is twofold. First, we wanted to observe if we could replicate the findings of Study 1 regarding the moral perceptions of social dilemmas. This is of crucial importance to social science and psychology, as previous work has documented a widespread problem in the replication of findings of several studies (see, for instance, Ioannidis, 2005, and more recently Camerer et al, 2018). Second, we wanted to investigate whether Morality As Cooperation theory (henceforth, MAC), which is rooted in the theory of nonzero sum games and hence closely related to social dilemmas, performed better than MFT in capturing individual differences in the moral judgments of subjects.

### 2.4.1. *Methods*

#### 2.4.1.1. *Procedures*

The only difference between the experimental design of Study 1 and Study 2 is that, in Study 2, we used the morality as cooperation questionnaire developed in Curry et al (2019) rather than the moral foundations questionnaire. We held constant, as far as possible, the other procedures of Study 1 so that the replication was as informative as possible.

The Morality As Cooperation Questionnaire was built to operationalize MAC, which was developed in Curry (2016) and subsequently discussed in Curry et al (2019), Gellner et al (2020), and Curry et al (2020 and 2021). Put shortly, MAC presents a theoretical framework that depicts morality as (i) based on biological and cultural adaptations to problems of cooperation; and (ii) based on seven types of cooperation related to the literature of evolutionary game theory.

To capture the seven types of cooperation that form the basis of morality, MAC proposes the foundations of *Family Values*, based on the cooperation problem of allocating resources to one's kin; *Group Loyalty*, based on coordination games; *Reciprocity*, based on social dilemmas; *Heroism* and *Deference*, based on hawk and dove games; *Fairness*, based on bargaining games; and *Property Rights*, based on interpersonal conflicts over resources[9]. For the remainder of the chapter, we shortened the names of the foundations to Family, Loyalty, Reciprocity, Heroism, Deference, Fairness, and Property so that, as the foundations of MFT, they can be referenced with the use of a single word.

[9] For the sake of consistency within the Study 1 and Study 2 sections, and at the risk of being imprecise, we also name the seven items of MAC as *foundations*. Curry et al (2019) use the word foundation only to refer to Moral Foundations Theory, whereas they refer to their items as *types of cooperation*. However, as their questionnaire uses the same elicitation tool as the one developed in Moral Foundations Theory, we found it convenient to name whatever is elicited with the same tool by the same name. Additionally, and for the remainder of the chapter we shortened the names of the foundations to Family, Loyalty, Reciprocity, Heroism, Deference, Fairness, and Property, so that, as the foundations of MFT, they can be referenced with the use of a single word.

Crucially, Curry et al's (2019) Moral Foundations Questionnaire elicits a score for each of these moral foundations, ranging from 0 to 100. As with the foundations elicited with the Moral Foundations Questionnaire, the score is to be interpreted as the importance of a given aspect for one's own conception of morality; a higher score denotes a higher relevance of a given aspect to one's own conception of morality. Analogously to Study 1, we use the standardised score of each foundation to test whether, and, if so, how, MAC's moral foundations influence the moral judgments of the scenarios of social dilemmas.

#### 2.4.1.2. *Participants*

To give MFT and MAC the same statistical power, we recruited the same number of participants ($n = 398$) for Study 2. As in Study 1, we recruited the subjects from MTurk. To minimise the difference in socio-demographic background of our participants between studies, we carried out the experiment at the same times and days of the week as in Study 1 (average age of participants: 36.53 years; 43% females; 51% liberals; and 40% religious). Our subjects capture the qualitative features regarding MAC-Q reported in Curry et al (2019) – See Appendix A.2.

### 2.4.2. *Results*

#### 2.4.2.1. *Moral Judgments*

We use the same main tool of analysis – the Moral Evaluation Functions of free riders, half, and full contributors – as for Study 1. Figure 2.5 plots the MEFs of Study 2.

Data from Study 2 successfully replicate the five main findings that emerged in Study 1, as Figure 2.5 shows[10]. First, the MEF's are substantially different from the moral neutrality benchmark. Second, our subjects perceive Person A to be more praiseworthy the higher his or her effective contribution, for any given effective

[10] We carry out a pairwise comparison of means between equivalent scenarios of Study 1 and Study 2. We observe no statistical difference in any of the scenarios, which we take as evidence that we also replicate quantitatively the findings of Study 1. Appendix A.6 provides a table with these pairwise comparison of means of impartial moral judgments of scenarios between studies.

contribution pattern of Persons B and C. Third, the shape of the MEF's is decreasing whenever $C_A \leq C_{BC}$ and flat otherwise. Fourth, our subjects perceive free riders as less blameworthy and half contributors as less praiseworthy when Person B and Person C act differently (i.e., $d = Unequal$). And fifth, the framing of the decision situation influences the moral judgments of our scenarios, especially of those where Person A effectively contributes 10.

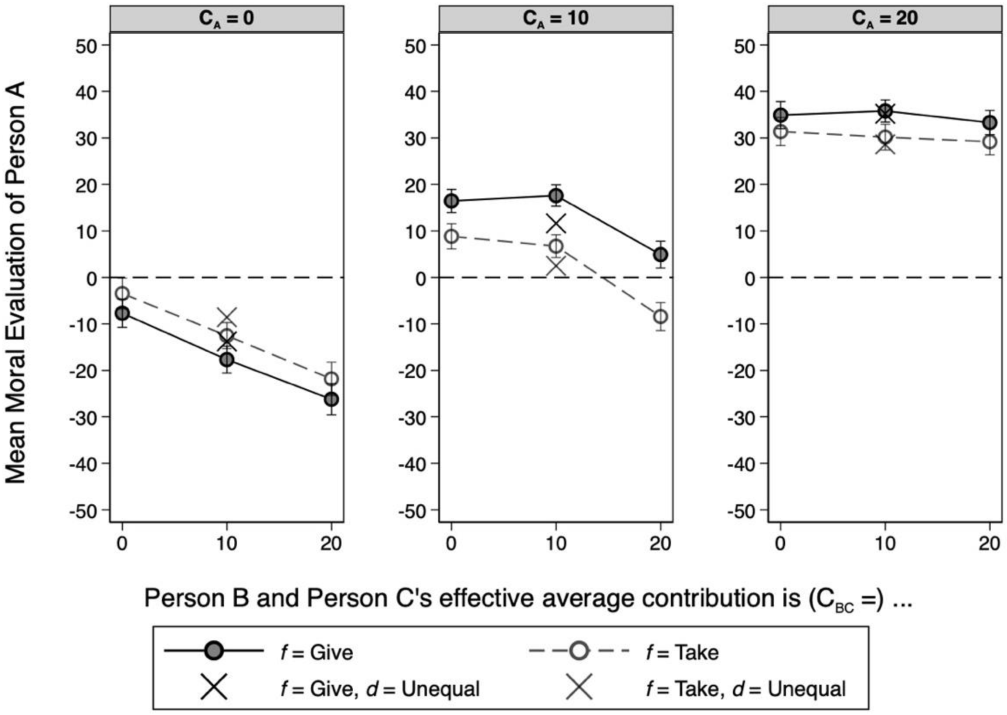


**Figure 2.5.** *Mean Moral Evaluations (of impartial spectators) ascribed to Person A - Study 2*

However, there are two minor differences between the data from both studies. First, the MEF's of $C_A = 0$ for the Give and Take treatments are now parallel, whereas in Study 1 the slope of the MEF of $C_A = 0$ was steeper for the Take treatments. Second, in Study 2 the MEF's of effective full contribution tend to vary according to the framing of the decision situation: full contributors are judged by our subjects, on average, to be more praiseworthy in the Give frame than in the Take frame of the social dilemma.

We additionally replicate the individual differences in subjects' moral judgments that we found in Study 1. We report the violin plots for the data of Study 2 in Figure

2.6. As in Study 1, the density of the moral judgments of scenarios spread over a substantial part of the range of the moral ratings scale. It is noteworthy to see that the dispersion in moral judgments is similar in Give and Take treatments.

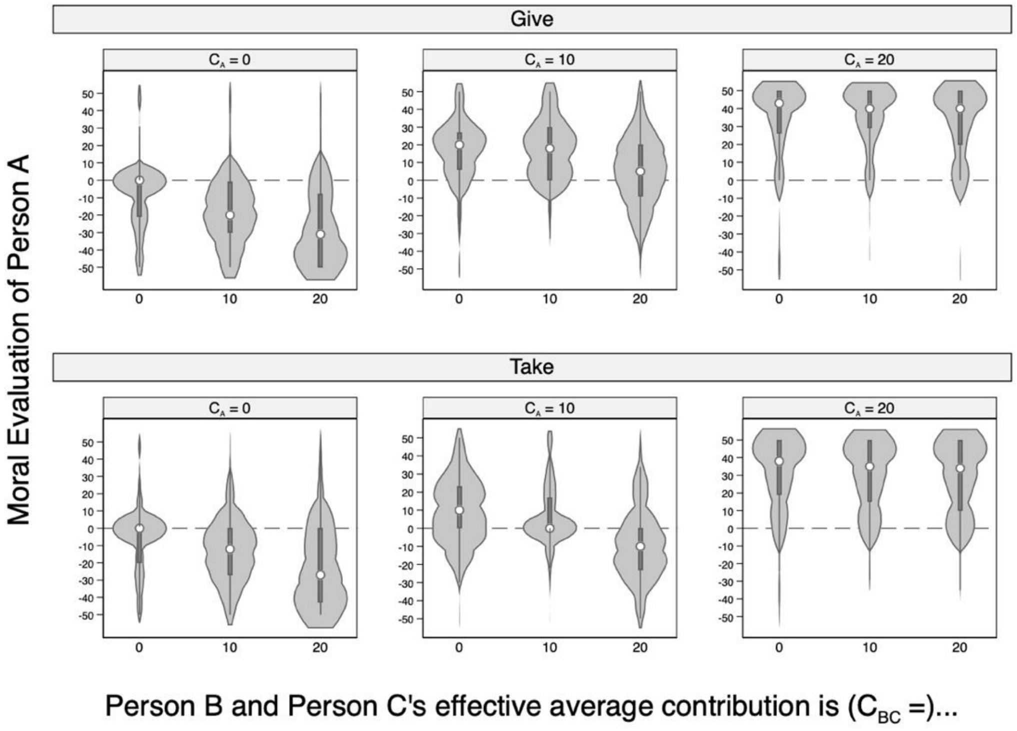


**Figure 2.6.** *Violin plots of the moral judgments of scenarios – Study 2*

We furthermore provide, in Table 2.3 and for all levels of $C_A$, the share of subjects displaying one of the five shapes of MEF's described earlier. As in Study 1, Table 2.3 documents the widespread adherence to different shapes of MEF at all levels of $C_A$, and a systematic adherence to different shapes of MEF's at different levels of $C_A$. Most remarkably, we replicate, both qualitatively and in size, the proportions of adherence of each shape of MEF at all levels of $C_A$. More specifically, some patterns emerge that match what we observed in the data of Study 1: (i) a decreasing MEF is most common when judging free riders, and its adherence rate is decreasing in the level of $C_A$; (ii) the increasing MEF is most common when judging full contributors and is increasing in the level of $C_A$; (iii) the adherence rate to a triangle MEF is most common at $C_A = 10$; and (iv) the adherence rate of a flat MEF is most common when judging full contributors, but is also substantially present when judging free riders.

Unlike in Study 1, this time we do find evidence of a framing effect. More specifically, we find that the distribution of adherence to different shapes of MEF's between frames is statistically significant at $C_A = 10$, being an adherence to a decreasing MEF at $C_A = 10$ substantially more common in the Take treatments.

**Table 2.3.** *Proportions of the different shapes of the MEF's of each level of $C_A$ for each frame – Study 2*

| | $C_A = 0$ | | $C_A = 10$ | | $C_A = 20$ | |
|---|---|---|---|---|---|---|
| | Give | Take | Give | Take | Give | Take |
| Flat | 22.66% | 19.49% | 17.73% | 17.44% | 34.48% | 34.87% |
| Decreasing | 55.17% | 50.26% | 33.00% | 46.67% | 23.65% | 22.56% |
| Increasing | 6.40% | 6.67% | 8.87% | 7.18% | 17.24% | 12.82% |
| Triangle | 9.85% | 15.90% | 31.53% | 24.10% | 11.33% | 11.79% |
| Reverse Triangle | 5.91% | 7.69% | 8.87% | 4.62% | 13.30% | 17.95% |
| Chi 2 test | $\chi^2_{(4)} = 4.24; p = 0.37$ | | $\chi^2_{(4)} = 9.64; p = 0.05$** | | $\chi^2_{(4)} = 2.74; p = 0.60$ | |

### 2.4.2.2. *Morality As Cooperation theory*

We follow the same approach to analyse the incidence of the moral foundations of MAC in the moral judgments of the MESD scenarios. That is, we estimate equation (2.1) for all the seven foundations, standardising the score of each foundation using the formula presented in equation (2.2). Furthermore, we present the data in figures, as we did for Study 1, and report the same statistical tests we reported in the previous section. Also as in the previous section, in the main text of the chapter we only report the results for the Reciprocity and the Fairness foundations, which are the ones more closely related to social dilemmas and relegate to appendix A.4.2. the discussion of the remaining five foundations. Strictly speaking, the Reciprocity foundation is the one that captures the cooperation problem of social dilemmas. However, the fairness foundation aims to capture the problem of bargaining and allocation of resources. As differential contributions have distributional consequences in a public goods game, we thought this foundation as also capturing, to a certain extent, an important qualitative feature of some of our scenarios (i.e., the distributional consequences emanating from the strategy combinations made by each of the group members).

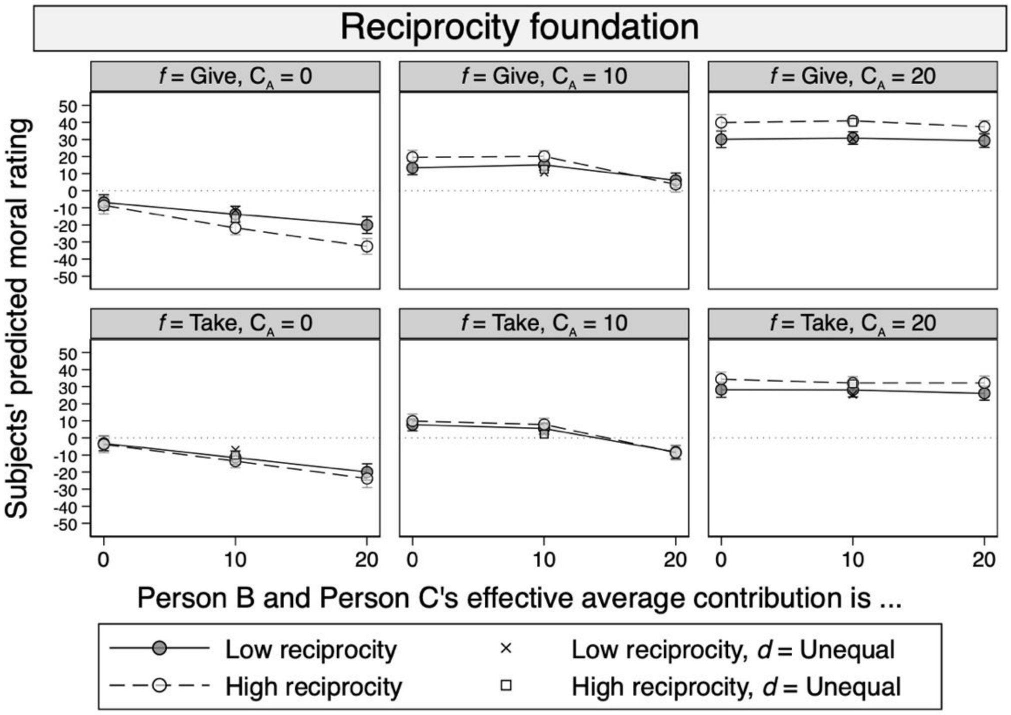


**Figure 2.7.** *Predicted MEF's evaluated at high and low scores of the Reciprocity foundation - Top Panels: Give scenarios; Bottom Panels: Take scenarios*

We start our discussion with the predicted moral judgments of high and low scores of the *Reciprocity foundation*, reported in Figure 2.7. This is the foundation of MAC that is most closely related to social dilemmas, so we expected reciprocity *ex ante* to be the one performing the best in driving moral judgments of MESD scenarios. When considering all 24 scenarios jointly, we find evidence of a significant effect of Harm on moral judgments ($F(24{,}397) = 2.51$, $p = 0.00$). More specifically, this effect is driven by the influence of Reciprocity in the moral judgments of the Give scenarios (Give frame: $F(12{,}397) = 2.72$, $p = 0.00$; Take frame: $F(12{,}397) = 1.03$; $p = 0.42$). Overall, a higher score in the Reciprocity foundation is associated with (i) more blameworthy judgments of free riders in the Give frame, driven by a more negative slope of the MEF of $C_A = 0$ (see top left panel); (ii) more praiseworthy judgments of half contributors when $C_A \geq C_{BC}$ in the Give frame (see top central panel); and (iii) more praiseworthy judgments of full contributors, especially in the Give frame (see top and bottom right panels). The effects are bigger than the ones associated with any of the foundations of MFT. Additionally, and in this time more in line with the previous findings regarding MFT, we do not find evidence of Reciprocity as a moderator of framing effects ($F(12{,}397) = 0.51,\ p = 0.91$). Nonetheless, there seems to be a tendency of a high score (relative to a lower score) in the Reciprocity foundation in generating (i) a harsher condemnation of free riders; and (ii) a more accentuated praise of half and full contributors.

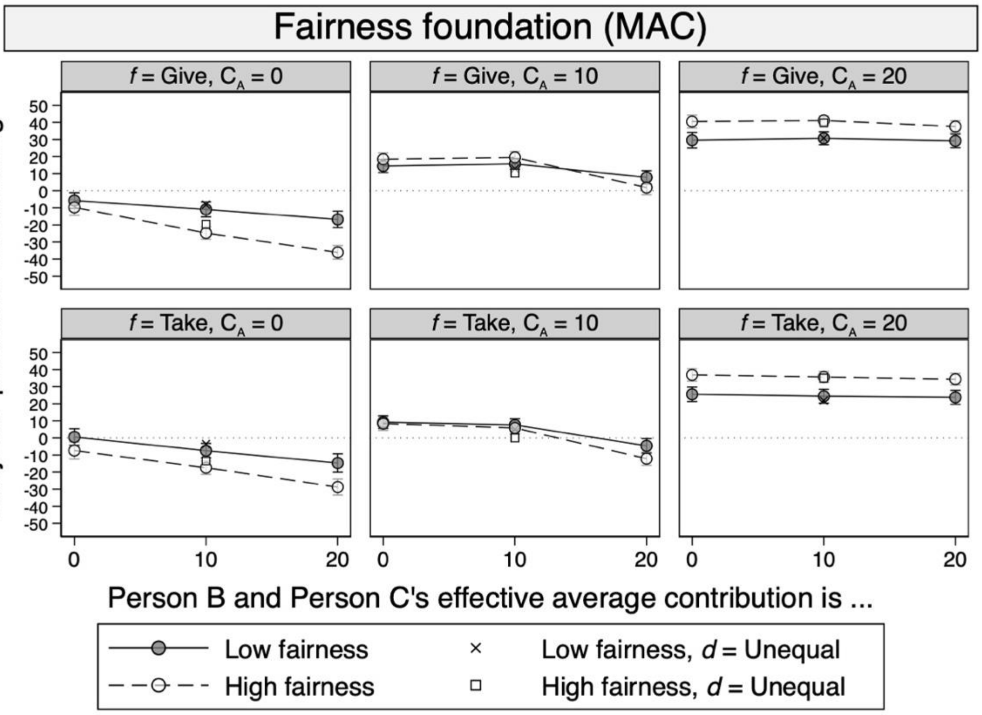


**Figure 2.8.** *Predicted MEF's evaluated at high and low scores of the Fairness (MAC) foundation – Top Panels: Give scenarios; Bottom Panels: Take scenarios*

Figure 2.8. reports the predicted moral judgments of all scenarios of high and low levels of the *Fairness foundation*. We do find strong evidence of a significant effect of Fairness on moral judgments; both at the aggregate level ($F(24{,}397) = 5.36, p = 0.00$) and at both Give and Take scenarios (Give frame: $F(12{,}397) = 5.86, p = 0.00$; Take frame: $F(12{,}397) = 4.86;\ p = 0.00$). As with the Reciprocity foundation, a higher score in the Fairness foundation is associated with (i) more blameworthy judgments of free riders in the Give frame, driven by a more negative slope of the MEF of $C_A = 0$ (see top left panel); and (ii) more praiseworthy judgments of full contributors in both the Give and the Take frames (see top and bottom right panels). Additionally, a high score in the Fairness foundation is also associated with (iii) more blameworthy judgments of free riders as well in the Take frame (see bottom left panel). In line with the findings of the Reciprocity foundation, we do not find evidence of Fairness being a moderator of framing effects ($F(12{,}397) = 0.96; p = 0.49$). That being said, there is a tendency of a high score (relative to a low score) in the Fairness foundation to generate a more accentuated praise of half contributors. Overall, the effects of the Fairness foundation are bigger than the effects of the Reciprocity foundation, especially the ones related to each foundation's moderating role of framing effects.

We can briefly summarise the results found from the other foundations as follows. First, all but the deference foundations are, overall, significant predictors of moral judgments. Second, the Group foundation have bigger effect sizes than the other remaining foundations. And third, some foundations seem better to explain the moral judgments of Give scenarios (i.e., Deference), others seem better to explain the moral judgments of Take scenarios (i.e., Property, Heroism, and Family), and only the Group foundation can predict moral judgments of Give and Take scenarios equally well.

### 2.4.3. *Discussion*

Overall, we replicate the main empirical patterns of moral judgments that we observed in Study 1 with striking precision. The effects of $C_A$, $C_{BC}$, $f$, and $d$ on the moral judgments of our subjects is qualitatively similar to the ones we found in Study 1. Furthermore, we also replicate the substantial individual differences in the impartial moral judgments of scenarios of social dilemmas, and we can use the shape of the MEF's to represent qualitatively the observed individual differences in moral judgments. Both the density of moral judgments and the adherence rate of different shapes of MEF's for all levels of $C_A$ match qualitatively and quantitatively with what we reported in Study 1, thereby making the results of this chapter stronger evidence of an underlying moral landscape of social dilemmas.

Lastly, individual differences in moral judgments are captured substantially better by MAC than by MFT, as evidenced by greater effect sizes and stronger significance in the effects. More specifically, (i) all but the Deference foundation have an overall significant effect on moral judgments; (ii) the Reciprocity and Deference foundations have a significant effect on the moral judgments of Give scenarios, the Property, Heroism, and Family foundations have a significant effect on the moral judgments of Take scenarios, and the Fairness and Group foundations have a significant effect on the moral judgments of both Give and Take scenarios; and (iii) the Reciprocity, Fairness and Group foundations are more significant contributors to the explanation of individual differences in the moral judgments of scenarios than the other foundations. However, like MFT, we do not find evidence of the Foundations in MAC as being driving forces behind the framing effects of the MEF's presented earlier, and the effect sizes observed are not able to capture the full extent of individual differences in moral judgments. Hence, further research is needed to document variables that can explain a greater share of the individual differences that we find.

## 2.5. Concluding remarks

The contribution of this paper is twofold. First, we contribute to economics by studying whether public goods provision, a type of social dilemma, is perceived as a moral issue. Second, we contribute to moral psychology by testing how well two recent theories aiming at explaining cooperativeness through a diverse array of moral foundations, those captured by Moral Foundations Theory and Morality As Cooperation theory, perform at shaping people's moral judgments of social dilemmas.

Amartya Sen has done extensive work on the relationship between ethics and economic behaviour, and on how the former could be a motivational force for the latter (see, especially, Sen, 1973 and 1977). Lately, Smith and Wilson (2017, 2019), based on Adam Smith's *Theory of Moral Sentiments*, proposed that moral concerns could drive prosocial actions. Although we remain silent about whether morality regulates behaviour in social dilemmas, in this chapter, as a first empirical step towards that goal, we explore the extent to which people consider several scenarios in social dilemmas as morally relevant. By measuring impartial moral judgments of both selfish and cooperative players of two social dilemmas, we provide a systematic picture of people's moral views of Give and Take social dilemmas. We show that social dilemmas are morally relevant decision situations. It is not only the negative side of morality, viz., that of free riders being perceived as blameworthy (shown by CDGK 2011), that matters; people also perceive contributors as praiseworthy. In sum, both blame for free riding and praise for cooperation are part of the moral landscape of social dilemmas. The next chapters will build on these results to analyse whether subjects' moral judgments of social dilemmas influence their unconditional contributions and contribution attitudes to public goods.

Moral Foundations Theory and Morality As Cooperation theory have an important aspect in common: the function of their moral foundations is to regulate selfishness and promote cooperativeness. The issue of moral motivation as a pathway to human behaviour has been extensively discussed, both in classical moral philosophy (see, e.g., Kant's Groundwork of the Metaphysics of Morals) and in modern moral psychology (see, for instance, Aquino and Reed, 2002). These two foundation-based theories provide a clear way to capture a wide range of topics that are perceived to be morally relevant, and that have the potential to explain the moral understanding, and regulation, of behaviour. In this chapter we put both theories to the test and contribute to the discussion of their validity by analysing whether the foundations of both theories can explain people's moral judgments of a canonical problem of cooperation: a social dilemma. Although both MFT and MAC provide foundations that can capture individual differences in moral judgments, the explanatory power of each foundation is rather small, and more research is needed to document which traits underly people's differences in their moral judgments of social dilemmas. Our data suggests, if anything, that people's moral worldviews as captured by MFT, and MAC play a small role when it comes to explaining moral judgments of free riding and cooperation. Furthermore, it is not only the specific foundations relevant to social dilemmas (i.e., the Fairness foundation of MFT and the Reciprocity foundation of MAC) that matter to explain our subjects' impartial moral judgments of our scenarios, revealing that different features inherent of social dilemmas (i.e., the distributional consequences of different combinations of actions) can trigger attitudes towards the moral perceptions of social dilemmas beyond what can be captured by the characteristic features of these decision situations (i.e., exploiting opportunities for mutual advantage).

To conclude, our findings shed some light into the moral dimension of social dilemmas. More specifically, people perceive free riders as blameworthy and contributors as praiseworthy. The framing of the decision situation, the contribution of the judged person, the heterogeneity in the non-judged group members' actions, and the effective average contribution level of the non-judged group members are important factors influencing our subjects' impartial moral judgments of social dilemmas. Furthermore, our results show the moral foundations of MFT, and MAC do not fully capture individual differences in moral judgments of problems of cooperation. Thus, further investigation needs to be carried out to determine what factors influence the subjective account of impartial morality that the data from our studies transpire.

# Chapter 3. The Morality of Voluntary Cooperation

## 3.1. Introduction

The main aim of this chapter is to explore the relationship between moral motivations and behaviour in public goods games. Four well documented empirical regularities motivate our present study. First, and contrary to the predictions of standard economic theory, people contribute positive amounts to a public good[11]. Second, people's contributions to the public good are conditional on the contribution of other group members (see Weimann, 1994, Keser and Van Winden, 2000 and Fischbacher et al, 2001 among others). Scholars interpret this finding as evidence that subjects' preferences for their contribution choices, henceforth *contribution preferences*, are conditional on the contributions of other group members. Third, beliefs play an important role in determining a subject's contribution to a public good (see, for instance, Fischbacher and Gächter, 2010). Finally, there are different types of public goods, and people's behaviour in each of them is different. Whereas some public goods do not exist in the first place and people need to provide them (*Provision problems*. e.g. universal primary education), others already exist, and people have to maintain them (*Maintenance problems*. e.g., the environment). Studies have shown that people tend to be more cooperative in Provision than in Maintenance problems[12]. These four facts have been replicated several times, but there is no agreement on their explanation.

The present chapter seeks to answer three different questions. Contribution preferences are assumed to exist and are used as a tool to predict contributions to a public good, paying little attention to the origin of those preferences. To shed some

[11] See Bohm (1972); Dawes et al (1977), Marwell and Ames (1979) and Isaac et al (1984) for early evidence, and Ledyard (1995) for a survey. See, more recently, Zelmer (2003), Chaudhuri (2011), and Gächter (2014).

[12] See Rutte et al (1987); Fleishman (1988); Van Dijk and Wilke (1997); Cookson (2000); Ellingsen et al (2012); and Gächter et al (2017). More recently, see Isler et al (2021).

light into their origin, we study whether moral motivations for avoiding blame and seeking praise are determinants of contribution preferences. We also study whether moral motivations influence a subject's contribution decision, either through this route or otherwise. Lastly, we study whether different moral views of what is blameworthy and praiseworthy in Provision and Maintenance problems can account for the different cooperativeness across the different problems[13].

Several reasons have been proposed to explain people's contributions to public goods: kindness, confusion, impure altruism, and conditional cooperation being the most notable ones[14]. Theoretical models of other-regarding preferences have been developed that can explain cooperation in public goods games[15]. Some of those models, such as reciprocity and guilt aversion, can additionally predict different cooperativeness across payoff-isomorphic Provision and Maintenance problems provided that subjects' beliefs vary across them (see Dufwenberg et al, 2011). These other-regarding preferences capture people's moral preferences indirectly. Yet, there is another explanation involving morality that has hitherto received little attention: that people's moral judgments of all the different choices in a public good problem drive their behaviour in those problems. Some economists propose people's moral impersonal considerations as a potential explanation of prosocial behaviour (see Harsanyi, 1955; Sen, 1977 and Smith and Wilson, 2019)[16]. In moral psychology and related disciplines it is common to define the purpose of moral judgments in terms such as "*to facilitate social regulation ..., ensuring cooperation among group members*" (Anderson et al, 2020, p.3)[17]. Building on both literatures, we present and

[13] Some initial evidence of different moral views of Provision and Maintenance problems is documented in Cubitt et al (2011) and chapter two of this thesis, where findings point out that free riding is seen as more reprehensible, and contribution as more praiseworthy, in Provision than in Maintenance problems.

[14] See, for instance, Andreoni (1988); Andreoni (1990); Weimann (1994); Andreoni (1995); Croson (1996); Palfrey and Prisbey (1996 and 1997); Anderson et al (1998); Keser and Van Winden (2000); Fischbacher et al (2001); Fischbacher and Gächter (2010); and Ferraro and Vossler (2010).

[15] See, e.g., Sugden (1984); Rabin (1993); Levine (1998); Fehr and Schmidt (1999); Bolton and Ockenfels (2000); Charness and Rabin (2002); Dufwenberg and Kirchsteiger (2005); Falk and Fischbacher (2006); Batigalli and Dufwenberg (2007); Cox et al (2007); López-Pérez (2008); Alger and Weibull (2013) and Masclet and Dickinson (2019).

[16] For other work in economics linking morality to prosocial behaviour, see, e.g., 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), Dal Bó and Dal Bó (2014), Hodgson (2014), Blasch and Ohndorf (2015), Hauge (2015), Daube and Ulph (2016), Capraro and Rand (2018) and Friedland and Cole (2019).

[17] There is an increasing literature on the relationship of morality and social regulation. See, e.g., Blasi (1984); Kohlberg and Candee (1984); Nucci (1996); De Waal (1997); Fischer and Ravizza (2000);

measure the role of two different procedures that capture two different moral motivations for action: praise seeking and blame avoidance.

To test the influence of those moral motives on contribution preferences and contributions, we present three experimental tasks to each of our subjects: the *M-experiment,* which measures – from an impartial spectator viewpoint – their moral judgments of all the different strategy combinations of the public good problem; the *P-experiment,* which allows us to measure contribution preferences; and the *C-experiment,* which measures contributions to a public good and beliefs about the contributions of the other group members. We frame the three experimental tasks in two different ways by eliciting moral judgments, contribution preferences, contributions and beliefs in Provision and Maintenance problems, manipulating those frames between subjects. To ensure we isolate the influence of the frame in subjects' choices, the Provision and Maintenance problems we study are isomorphic in terms of their payoff consequences. Using the moral judgments from the *M-experiment*, we generate predictions of contributions and contribution preferences for each subject, both for blame avoidance and praise seeking motives. We compare these predictions with subjects' actual choices in the *P*- and *C-experiments* to study the influence of moral motivations on behaviour in Provision and Maintenance problems.

Our results point to four conclusions. First, a moral motivation to avoid being blameworthy explains contribution preferences even when controlling for strong reciprocity. Second, moral motivations to avoid being blameworthy and to seek praise explain contributions even when controlling for contribution preferences. Third, contribution preferences are the main motivation behind people's contributions, but a big part of its success is due to the effect that moral motivations (esp. blame avoidance) have in shaping people's contribution preferences. Fourth, the logic behind contributions is different in Provision and Maintenance problems. Whereas the Provision of public goods is mainly explained by contribution preferences, the Maintenance of public goods is achieved, to a greater extent, by the direct effect of people's moral motivations in shaping their cooperation.

Our results shed some light on three important issues in the literature on cooperation. First, we find that moral motivations explain people's contribution

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); and Schein and Gray (2018).

preferences even when controlling for the contribution of the other group member. We show some evidence suggesting that, on top of such conditionality, people's choices of their contribution preferences tend to rely on what they believe is morally right or wrong.

Second, moral motives (avoid being blameworthy and seek being praiseworthy) explain why contribution choices sometimes deviate from people's contribution preferences. Whether this deviation is better captured by an additional, different preference for moral actions, or by moral actions that may not rely on preferences but, rather, on moral principles and rules as Sen (1977) and Smith and Wilson (2019) suggest is an open debate for future research to explore.

Third, we find that moral motivations are a better predictor of framing effects in contribution choices than other theoretical alternatives. Distributional preferences cannot predict different attitudes in payoff isomorphic games, and the alternative explanation is that psychological game theory models like reciprocity can predict a framing effect through different beliefs across frames (e.g., Dufwenberg et al. 2011). Even when controlling for beliefs, the main difference in contributions across frames is due to our two moral motives as the different beliefs we observe would predict a pattern of behaviour opposite to the one observed in the data. A higher belief in contributions of the Provision frame would suggest a higher predicted contribution in Provision frames, but the average contribution across frames is not significantly different. A higher reliance on praise seeking and blame avoidance appear to be the motives sustaining contributions in the Maintenance of public goods.

The chapter is organised as follows. Section 3.2 presents the experimental design. Section 3.3 presents the theoretical predictions, focusing on presenting the different procedures we use to generate predictions of contributions and contribution preferences. Section 3.4 discusses the results and Section 3.5 concludes.

## 3.2. Experimental design

The experiment consists of two different parts: experimental tasks and three questionnaires. The experimental tasks were based on the same decision situation: a linear, one-shot, simultaneous move, two-person public goods game. The basic setup was framed either as a give-some or a take-some problem, which we refer to respectively as Provision and Maintenance problem. We manipulated the frame between subjects.

In the *Provision* problem, each group member is endowed with 30 tokens. He or she has to decide how many to *give* to the public good, which we referred to within the experiments as a “group project”. Each group member’s possible levels of *giving* were 0, 10, 20 and 30. The material payoff function of subject $i$ in the Provision problem is:

$$\text{(3.1)} \qquad \pi_i = 30 - g_i + 0.75 \times \left(\sum_{j=1}^{2} g_j\right)$$

where $g_i$ refers to the number of tokens a group member *gives* to the public good. Per token kept for him or herself, subject $i$ gets one point. The marginal per capita return of the public good (henceforth, MPCR) is 0.75, meaning that subject $i$ receives 0.75 points per token given to the public good regardless of who gave it.

In the *Maintenance* problem, the public good is initially populated with 60 tokens. Each group member has control over 30 of them. Their task is to decide how many, up to 30, to *take* from the public good. Each group member’s possible levels of *taking* were 0, 10, 20 and 30. The material payoff function of subject $i$ in the Maintenance problem is:

$$\text{(3.2)} \qquad \pi_i = t_i + 0.75 \times \left(60 - \sum_{j=1}^{2} t_j\right)$$

where $t_j$ refers to the number of tokens a group member *takes* from the public good. Each token taken from the public good gets subject $i$ one point. The MPCR was the same as in the Provision problem, 0.75, meaning that per token maintained in the public good subject $i$ receives 0.75 points regardless of who decided to leave it there. In both decision situations, the total points a subject gets from the actions of all group members are converted to monetary payoffs by dividing the number of points by 35.

The Provision and Maintenance problems are isomorphic in terms of monetary consequences. Exploiting this fact, we define the term *effective contribution* of subject

$i, c_i = g_i = 30 - t_i,$ as the strategy that yields the same payoff in both frames. It then follows that there are no differences in terms of equilibrium play in both frames: standard theory predicts $c_i = 0$ in both frames.

### 3.2.1. *Experimental tasks*

We use three different experimental tasks in our experiment, which we refer to as the *M-experiment,* the *P-experiment,* and the *C-experiment*. The *M-experiment* consists of eliciting the moral judgments that subjects have, from an impartial spectator viewpoint, of all the strategy combinations of the decision situation presented to them. The *P-experiment* elicits subjects' unconditional effective contributions and effective contributions conditioned on what the other group member does. The *C-experiment* elicits subjects' unconditional effective contributions and beliefs about the play of their partner within the setup of our decision situation.

#### 3.2.1.1. *The M-experiment*

Moral evaluations of social dilemmas were first studied in Cubitt et al (2011). The structure of the *M-experiment* follows their (and the previous chapter's) implementation, with some amendments. In short, subjects are first presented the public goods game in a given frame. The group members are referred to as Person A and Person B. Subjects are asked to rate the morality of Person A from an *impartial spectator* viewpoint in different scenarios: third parties that are not involved in the decision situation at the moment of judging. This is akin to how some theories of moral philosophy envisage moral judgments as drivers of behaviour (for an in-depth coverage of the subject see, for instance, the articles in the issue of *Ethics*, 101(4); Mendus, 2002; Raphael, 2009; and Feltham and Cottingham, 2010). This view has received some support in economics, especially in the work of Konow (2009, 2012).

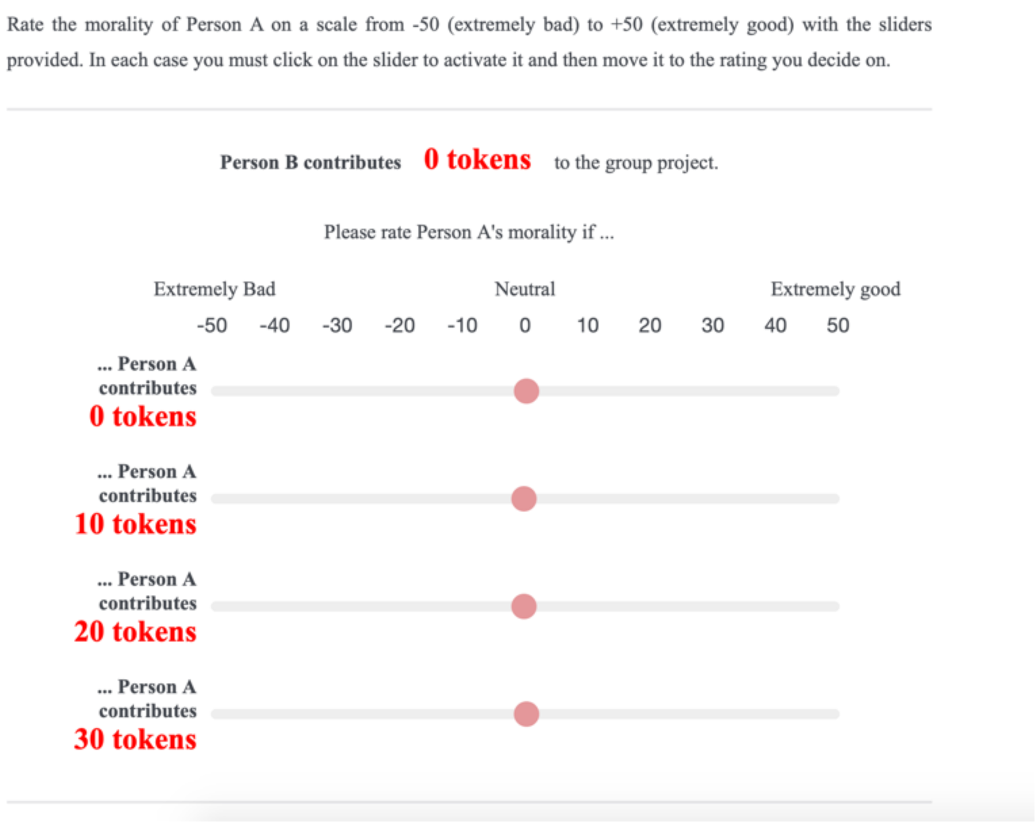
The *M-experiment* asks the subjects to rate the morality of person A in several scenarios. The moral rating ascribed to Person A in a given scenario is a number on a scale from -50 to +50, were -50 is labelled as extremely bad, +50 as extremely good and 0 as morally neutral. We introduce some notation in order to define the scenarios in a compact way. Let $C_A := \{0,10,20,30\}$ – with typical element $c_a$ – be the set of possible effective contributions of Person A, $C_B := \{0,10,20,30\}$ – with typical element $c_b$ – be the set of possible effective contributions of Person B and $F := \{provision, maintenance\}$ – with typical element $f$ – be the set of the frames of the decision situation we study. The set of scenarios of the *M-experiment* comprises all ordered triples within the Cartesian product $C_A \times C_B \times F$, with typical element $x \coloneqq \langle c_a, c_b, f \rangle$; where each element $x$ defines each scenario. Thus, the *M-experiment* comprises 32 scenarios, 16 per frame. We presented the 16 scenarios to each subject for the frame they were randomly allocated to (within-subjects design), which allowed us to elicit the moral perception of our subjects about person A for every strategy combination of the relevant decision problem. This entails a mixture of a between-subjects design, given that the frame was manipulated between subjects, and a within-subjects design (see above). Figure 3.1 presents an example screenshot of how we presented the scenarios to subjects facing the Provision problem.

There are three aspects of the *M-experiment* that are noteworthy. First, we did not link moral judgments to monetary incentives as the very nature of moral judgments, as evidenced by the data of the previous chapter, is inherently subjective. Second, the way we elicit moral judgments in the *M-experiment*, and how we subsequently use them, is a departure from the work of Smith and Wilson (2017, 2019), who draw some general principles from their interpretation of Adam Smith's Theory of Moral Sentiments[18]. The philosophical doctrine of moral sentimentalism endorses the individualism of people's normative principles (see, for instance, Schmitter, 2020)[19]. Indeed, our data and the data of the previous chapter corroborates this stance: subjects' moral perceptions of free riding and positive contributions in social dilemmas are importantly heterogeneous. This can be seen as an empirical challenge to the universalisation of moral views. Given such heterogeneity, it is natural to let each subject express their own moral views of the decision situation without imposing any more structure on them than is implicit in the procedures described when presenting

[18] Smith and Wilson (2017) derive some general ethical principles of behaviour from Adam Smith's Theory of Moral Sentiments and compare subjects' behaviour with those general principles. In contrast, we measure each subject's own moral views about the decision situation and use them to explain their play in the decision situation.

[19] For instance, David Hume, in his essay Of the Standard of Taste (1757, ST 2 and 3), stated that 'The sentiments of men often differ with regard to beauty and deformity of all kinds … 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 conduct and manners, the difference among men is really greater than at first sight it appears'.

the *M-experiment*. Finally, we opt for measuring the morality of persons performing actions rather than the morality of actions themselves as the existing literature suggests blame can only be assigned to the former, and our goal is to test blame and praise as moral motivations for action (see Malle, Guglielmo and Monroe, 2014 for a discussion).



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**

**Figure 3.1**.*Screenshot of some scenarios of the M-experiment (Provision problem)*

### 3.2.1.2. *The P-experiment*

The *P-experiment* was first presented in Fischbacher et al (2001) and was subsequently named the *P-experiment* in Fischbacher and Gächter (2010). The validation of the method was investigated in Fischbacher et al (2012). The *P-experiment* consists of two tasks: an *unconditional choice* and a *conditional choice* task. Subjects had to answer both tasks without knowing the answers of other group members. In the *unconditional choice* task, a group member decides his or her *effective contribution*. The *conditional choice* task uses a modified version of the strategy method, developed by Selten (1967), whereby subjects have to state their preferred effective contribution as a function of each of the potential effective contributions of the other group member. We refer to this vector of effective contributions as a subject's *effective contribution preferences*.

To ensure incentive compatibility, each task is chosen with a probability of 50% for payment, not disclosing the chosen task for payment until the experiment had ended. This meant that both tasks were payoff-relevant, as they could potentially be chosen for payment. As subjects are randomly matched in groups of two, in practice this meant that for one subject in the group the unconditional choice task was chosen for payment while for the other group member the conditional choice task was chosen for payment.

### 3.2.1.3. *The C-experiment*

The *C-experiment* consists of two tasks: an *unconditional choice* task and an incentivised *belief elicitation* task. Subjects were matched with different people in the *P-* and *C-experiments* in order to avoid any strategic considerations or learning influencing the decisions of the *C-experiment*[20]. In the *unconditional choice* task subjects had to decide their *effective contribution* without knowing the choice of the other group member. In the *belief elicitation* task subjects had to guess the other group member's choice in the *unconditional choice* task of the *C-experiment*. Each subject earned $0.34 more if their guess was correct and nothing otherwise[21].

We opted for a single round of unconditional choices instead of the 10 rounds in Fischbacher and Gächter (2010) because, unlike Fischbacher and Gächter (2010), we are not interested in how cooperation develops over time. Additionally, we wanted to keep the set of scenarios in the *M-experiment* described above manageable, and the time component substantially increased the set of feasible strategies for each player.

[20] To ensure that no learning could take place, subjects did not learn the choices of his previous partner in the *P-experiment* before engaging with the *C-experiment*.

[21] The choice of the incentivisation rule is different from previous ones used in the literature (see Croson, 2000 and Gächter and Renner, 2010). The choice of the incentivisation mechanism was done to prevent any strategic motivations biasing the beliefs. Under the previous incentivisation schemes and given that a guess of 10 or 20 has a maximum mistake of 20 whereas a guess of 0 or 30 has a maximum mistake of 30, subjects may be induced to report partial effective contributions as their guesses to maximise their minimum earnings from the guessing task. The smaller strategy set we employ made this more salient to the subjects

### 3.2.1.4. *The Questionnaires*

Subjects answered three different questionnaires after they finished the *M-experiment* task: the Moral Foundations Questionnaire developed by Graham et al (2011. Henceforth, MFQ), the Morality As Cooperation Questionnaire developed by Curry et al (2019. Henceforth, MAC-Q) and a sociodemographic questionnaire to gather some information on the background characteristics of our participants. The moral questionnaires capture people's general moral worldviews. The theories behind the questionnaires posit that morality's purpose is to regulate behaviour, and that morality is based on several foundations that are intuitive and built in human nature prior to experience. Foundations in MFQ are measured on a scale ranging from 0 to 5 whereas foundations in MAC-Q are measured on a scale ranging from 0 to 100. For an in-depth coverage of Moral Foundations Theory (henceforth, MFT), see Haidt and Joseph (2004, 2007), and Haidt and Graham (2009). For a coverage of the theory behind Morality As Cooperation theory (MAC), see Curry (2016). For the original papers presenting MFQ and MAC-Q and showing their internal and external consistency, see Graham et al (2011) and Curry et al (2019) respectively.

### 3.2.2. *Treatments*

In all treatments, subjects were first presented with a description of the public goods problem. The presentation of the task did not refer to them as one of the members of the group as the *M-experiment* was to be done as a third party. The subjects were explicitly briefed on when they were a part of the group (in the *P-* and *C-experiments*) and when they were making judgments as an outside, impartial observer (in the *M-experiment*). Immediately after presenting them with the corresponding decision problem, we asked subjects to answer 10 control questions to make sure they understood the monetary consequences of five strategy combinations for the two group members. Subjects were allowed to continue with the experiment once they answered the control questions correctly. We told them that the experiment consisted of several parts but did not disclose the content of each part until it was reached. This procedure minimises the danger of spillover effects between tasks.

The frame of the decision situation was manipulated between-subjects. Subjects were randomly assigned to the Maintenance (Provision) frame and took part in the Provision (Maintenance) version of the *M-*, *P-* and *C-experiments*. All the other tasks (the three questionnaires) were the same for all the subjects.

To control for order effects, we manipulated between-subjects the sequence in which the tasks were presented to subjects. To make the number of potential orders manageable, we constrained the different orders in which all the tasks could be presented according to the following set of criteria. First, the *M-experiment* was always followed by the sociodemographic questionnaire. Second, the sociodemographic questionnaire was always followed by the two moral questionnaires (MFQ and MAC-Q). Third, the *P-experiment* was always presented right before the *C-experiment*. Subject to these constraints, we manipulated (i) the order of the two moral questionnaires (the MFQ was presented before or after MAC-Q); (ii) the order of the two tasks within the *P-experiment* (the *unconditional choice* task was presented before or after the *conditional choice* task); (iii) whether the *P-experiment* and the *C-experiment* were presented before the *M-experiment* or after the two moral questionnaires. This generated eight potential different sequences in which to present all the tasks to the subjects. Together with the frame manipulation, this generated 16 different treatments. Subjects were randomly allocated to one of these 16 treatments.

### 3.2.3. *Participants and procedures*

We ran the experiment online on MTurk with 603 participants: 324 in the Provision treatments, and 279 in the Maintenance treatments. The average age was 35.3 years; 48.1% were female; 67.7% identified as liberals and 52.3% said they are religious[22].

The procedures were as follows. Participants were only able to participate if they had not participated in the experiments of the previous chapter. Subjects were paid a participation fee of $3 provided they completed the whole experiment. On top of that, they were paid according to their answers in the *P-experiment* and the *C-experiment*.

[22] In the sociodemographic questionnaire, we asked for self-reports of political ideology and religiousness. Regarding the political ideology scale, subjects to the left (right) were classified as liberals (conservatives). Regarding the religiousness scale, subjects that answered “not at all religious” were classified as not religious. Otherwise, we classified them as religious. For more information on all the questions asked in the sociodemographic questionnaire and the scaling of each question, one should refer to the experimental instructions presented in Appendix B.

The consequences of the decisions in the *P-experiment* and the *C-experiment* were framed as experimental points, that were subsequently converted into dollars. The average payment was $5.28 (including the $3 participation fee), and the average completion time was 28.5 minutes (average payment rate: $11.1/hr). Within the *P-* and *C-experiments,* subjects did not receive any feedback on their performance on the previous tasks to prevent any spillover effects between tasks.

## 3.3. Theoretical predictions

In this section we present three different procedures: ABC, blame avoidance and praise seeking. Each procedure generates, for each subject, a point prediction of his or her *effective contribution* in the *C-experiment*. Additionally, blame avoidance and praise seeking generate, for each subject, a prediction of his or her schedule of effective *contributions* in the *P-experiment*. We briefly describe these procedures and, for the key two innovations – blame avoidance and praise seeking – provide examples to show how they generate their predictions.

### 3.3.1. *The ABC approach*

The ABC approach explains cooperation as a function of a subject's attitudes and beliefs. The ABC approach was developed in Fischbacher and Gächter (2010), Fischbacher et al. (2012) and Gächter et al (2017) who also introduced the ABC terminology. ABC uses *effective contribution preferences* and beliefs to make a point prediction of a subject's effective contribution in the *C-experiment*. The experimenter elicits a function $a_i: C_{-i} \rightarrow C_i$ for subject $i$ using the *conditional choice* task of the *P-experiment*, where $C_i$ and $C_{-i}$, with typical elements $c_i$ and $c_{-i}$, are, respectively, the sets of the effective contributions of subject $i$ and the other group member in the decision situation. The function $a_i$ reveals the preferred effective contribution of subject $i$ for each effective contribution of the other group member, hence capturing subject $i$'s *effective contribution preferences*. The *belief elicitation* task reveals the belief that subject $i$ holds about the effective contribution of the other group member in the *C-experiment*. The strategy method assumes $a_i$, subject $i$'s *effective contribution preferences*, as governing subject $i$'s decisions and $b_i$ as an input to such preferences in order to generate a point prediction about $i$'s effective contribution in the *C-experiment*: $a_i(b_i) = \hat{c}_i \in C_i$.

### 3.3.2. *Blame Avoidance*

The moral psychology literature has proposed moral blame to be a regulator of the social world (See Darley and Shultz, 1990; Cushman, 2013, Malle et al, 2014 and Malle, 2021). More recently, some economists have proposed "not wanting blame and to not be blameworthy" as a "human motivation for action" (Smith and Wilson, 2019, p. 77)[23]. These two strands of the literature compelled us to study whether a moral motivation to avoid being blameworthy is a driver of people's behaviour in public good problems. We developed (and present below) the *Blame Avoidance* procedure (henceforth, BA) as a technique to generate, for each subject, predictions of their effective contribution preferences and effective contributions.

We first introduce some terminology that we use when describing the BA procedure. We define a given strategy as *blameworthy* for subject $i$ when subject $i$'s moral rating of a person performing that strategy is lower than 0, neutral when $i$'s moral rating of a person performing that strategy is equal to 0 and praiseworthy otherwise. BA is constructed by applying three different concepts: the desire to avoid being blameworthy and the assumptions of *self-love* and *self-command*. The desire to avoid being blameworthy implies that, among the feasible strategies for a given subject, he or she will refrain from choosing those which he or she considers would make him to be blameworthy. The assumption of *self-love* is similar to Smith and Wilson's (2019, P.69) axiom 0 and implies that a subject is inclined to choose the strategy which maximises his or her own material payoff. The assumption of *self-command* is also similar to Smith and Wilson (2019, pp.101-103), and follows from the impartial spectator design feature of the *M-experiment*: a subject deciding whether to behave according to his or her moral views consults the moral views that he or she

[23] For a different approach to modelling blame within economics, see Çelen et al (2017). However, their approach is fundamentally different from the one proposed by Smith and Wilson (2019) and the one we present here, as their conception of *blame-freeness* relies on a self-centred perspective based on what one would do in the situation of the other. This is, perhaps, more akin to Scanlon's (1998) *contractualism*, or to Darwall's (2006) *second-person standpoint* within moral philosophy, rather than to our strict focus on an impartial spectator stance.

believes an impartial spectator – that is, a disinterested actor in the scenario judged of – should hold.

The BA procedure applies these three different concepts in sequence to obtain predictions of our subjects' behaviour. First, the desire to avoid being blameworthy and the assumption of *self-command* restrict the strategies a subject will consider doing to be those for which him or her judged Person A as morally neutral or praiseworthy in the *M-experiment*. The key point to note here is that *self-command* implies that a subject will consider whether he or she is blameworthy whenever he or she considered, in the *M-experiment*, Person A to be blameworthy. Lastly, the assumption of *self-love* chooses, from this subset of strategies, the one that gives him or her the highest material payoff.

As an example of how BA makes point predictions from the moral judgments we elicit in the *M-experiment*, let's consider we present the Provision problem to subject $i$ and his or her belief is $b_i = 20$. If subject $i$ believes that the other group member will effectively contribute 20, then $b_i$ together with *self-command* will restrict the set of relevant moral judgments to those he or she expressed in scenarios where $c_b = 20$. There are two reasons for this. First, *self-command* imposes that the relevant moral judgments for subject $i$ are those he or she expressed in the *M-experiment*. Second, $b_i$ implies that subject $i$'s strategies will be embedded in scenarios where the other group member chooses $b_i = 20$. As Person B is the other group member in the scenarios of the *M-experiment*, *self-command* and $b_i = 20$ restrict the relevant moral judgments to be those of the scenarios where $c_b = 20$. Hence, those of the form $\langle c_a, c_b = 20, provision \rangle$. Let's assume subject $i$'s vector of the moral ratings of those scenarios is (-20, 0, +20 +20), where the first (resp. second; resp. third; resp. fourth) element refers to the scenario where $c_a = 0$ (resp. $c_a = 10$; resp. $c_a = 20$; resp. $c_a = 30$). From the four strategies available to Person A, only free riding makes Person A to be perceived as blameworthy (-20 < 0) in the *M-experiment*. And, by *self-command*, it follows that subject $i$ will also perceive him or herself as being blameworthy when free riding. Hence, subject $i$ will refrain from free riding and will consider choosing any of the other strategies. The assumption of *self-love* induces subject $i$ to choose an effective contribution of 10, as it is the remaining strategy that gives him or her the highest material payoff. Hence, the BA procedure predicts that, given $b_i = 20$, subject $i$ will contribute 10 in the *C-experiment*.

Additionally, BA allows us to get predicted effective contributions for each effective contribution level of the other group member in both Maintenance and Provision games. The logic behind these predictions is the same as the one behind the prediction of the effective contribution in the *C-experiment*, which we have explained in detail above. The only difference is that instead of using $b_i$ to constrain the relevant moral judgments to those of scenarios where $c_b = b_i$, we use the effective contribution of the other group member, $c_{-i}$, to constrain the moral judgments to those of scenarios where $c_b = c_{-i}$. As there are four possible effective contributions of the other group member, BA will make four predictions, one per each effective contribution level of the other group member. The vector of these predictions is the *predicted effective contribution preferences* that BA makes. By comparing these predictions with subject $i$'s answers to each entry of the *conditional choice* task in the *P-experiment*, we can study whether BA is a good predictor of subject $i$'s effective contribution preferences.

### 3.3.3. *Praise Seeking*

In addition to blame as a motivation for social regulation, some recent research suggests that *moral praise* can act as a reinforcement of moral behaviour (Anderson et al, 2020). Although the literature on praise is relatively small when compared to that of blame, some authors have established an asymmetry between judgments of blame and praise, where, empirically, judgments of praise are not necessarily the opposite of judgments of blame (see Pizarro et al, 2003). Furthermore, some studies show that the tendency to blame wrong actions is orthogonal to the tendency to praise right actions (see Wiltermuth et al, 2010). Taking into account these considerations, we see as plausible that praise has a different role than blame in driving behaviour. We develop the Praise Seeking procedure (henceforth, PS) as an attempt to capture a moral motivation for doing what is perceived as most praiseworthy. PS is based not only on the moral psychology literature cited above, but also on Smith and Wilson (2019), who state that "human motivation for action arises from wanting praise and to be praiseworthy" (p. 77).

PS is constructed by applying three different concepts: the desire of being as praiseworthy as possible and the assumptions of *self-love* and *self-command*. Being as praiseworthy as possible implies that, among the feasible strategies for a given subject, he or she will only consider choosing those actions which he or she considers make him or her most praiseworthy. These three concepts are applied sequentially in a similar manner as in BA. First, the desire of being most praiseworthy together with *self-command* restricts the set of strategies a subject will consider choosing to those for which him or her rated Person A as most praiseworthy in the *M-experiment*. The assumption of *self-love* then chooses, from this subset of strategies, the one that gives him or her the highest material payoff.

Following the example above, let's assume subject $i$ is presented with the Provision problem and has $b_i = 20$. As in BA, and following the same logic, $b_i = 20$ will constrain the relevant moral ratings to be those where $c_b = 20$. Let's assume, for the sake of simplicity, that the vector of relevant moral ratings is the same one as above: (-20, 0, +20 +20). The strategies that make subject $i$ to perceive Person A as most praiseworthy are the effective contributions of 20 and 30, as +20 is the highest moral rating attached to Person A for scenarios where $c_b = 20$. *Self-command* makes subject $i$ to perceive him or herself as most praiseworthy when he or she effectively contributes 20 or 30 to the public good. Among those strategies, *self-love* induces subject $i$ to choose an effective contribution of 20 as it is the strategy that gives subject $i$ the highest material payoff. Hence, it follows that PS will predict subject $i$ to effectively contribute 20 in the *C-experiment*.

Also, as in BA, PS additionally gives us a prediction for each effective contribution level of the other group member. The vector of those predictions is the PS prediction for subject $i$'s *effective contribution preferences*. Again, by comparing these predictions with subject $i$'s actual behaviour in the *conditional choice* task of the *P-experiment* we can investigate whether PS is a driver of contribution preferences.

## 3.4. Results

### 3.4.1. *Moral views of Provision and Maintenance problems*

We plot the average moral judgments (with 95% confidence intervals) of all the scenarios of the *M-experiment* in Figure 3.2, dividing them into 4 different panels. Each panel contains all the average moral judgments referring to a given effective contribution level of Person A (the person being judged). The horizontal axis denotes the effective contribution level of Person B (the non-judged person), and the vertical axis denotes the moral rating of a given scenario, ranging from -50 (extremely bad) to +50 (extremely good). A moral rating of 0 implies that a given scenario has no moral significance and, as a *not-morally-relevant* benchmark, we plot a dashed horizontal line at that level in the four panels. Based on Cubitt et al (2011) and on the previous chapter of the thesis, we define the *Moral Evaluation Function of* $x$ (henceforth, MEF of $x$) as the average moral evaluation ascribed to the judged Person when he or she effectively contributes $x$, expressed as a function of the effective contribution of the non-judged person.

Each of the four panels in Figure 3.2 contains two MEF's, the one of Provision and the one of Maintenance problems. The MEF's of Provision problems are plotted as black, solid lines and the MEF's of Maintenance problems are plotted as black, dashed lines. The effective contribution of the judged Person is written at the top of each panel.

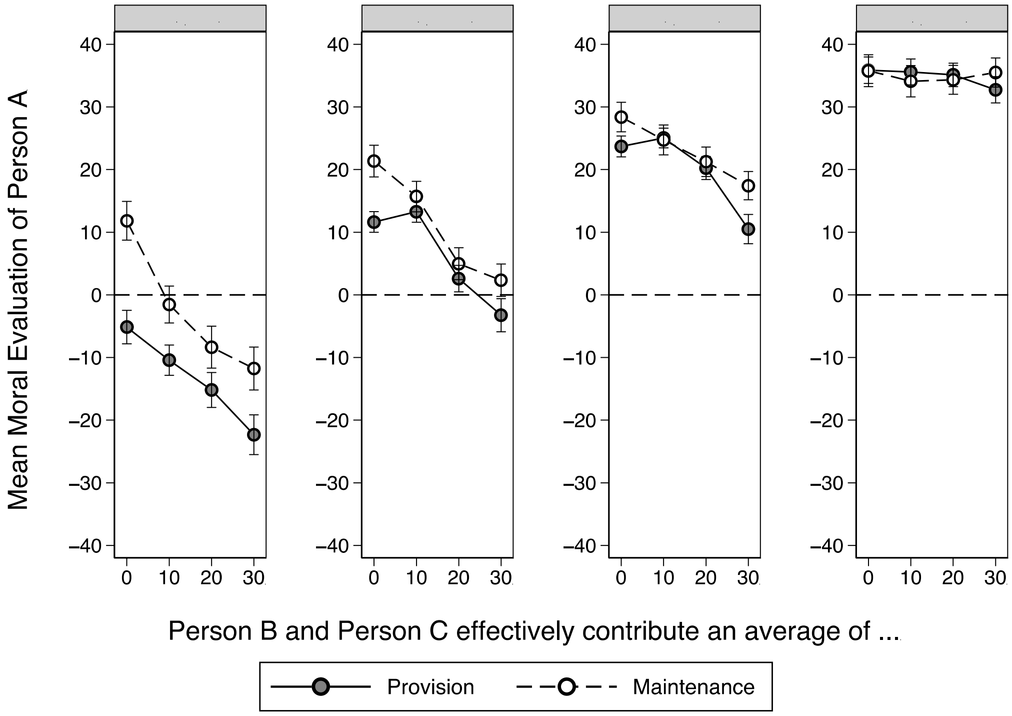


**Figure 3.2.** *Moral Evaluation Functions of all Effective Contribution Levels of the Decision Situation*

Four characteristics of the moral views elicited in the *M-experiment* are worth mentioning and replicate most of the qualitative findings of the previous chapter and of Cubitt et al (2011) for free riding. First, most of the scenarios have an average moral rating different from 0, evidencing that subjects do perceive Provision and Maintenance problems as being morally significant. Second, the morality ascribed to Person A is increasing in his or her own effective contribution. This is evidenced in Figure 3.2 by an upward shift in the MEF's of Person A the higher his or her effective contribution, being the MEF's of free riding the ones with the lowest moral ratings and the MEF's of full effective contribution the ones with the highest moral ratings. Third, the MEF's are decreasing in the effective contribution of Person B. This effect is moderated by the effective contribution of Person A: the negative slope of the MEF's is greatest in free riding and non-existing in effective full contribution. Finally, framing effects are mainly present at free riding: Person A is seen as more reprehensible when not giving than when taking everything. The MEF's of effective positive contributions do not vary substantially with the frame of the decision problem[24].

### 3.4.2. *Effective contribution preferences of Provision and Maintenance problems*

We follow Fischbacher et al (2001) in that we analyse effective contribution preferences by splitting our sample according to different contribution types. However, we use the refined classification scheme of Thöni and Volk (2018), which considers five different contribution types: Free riders, conditional co-operators, unconditional co-operators, triangle contributors and not classified (n.c.). Table 3.1 below shows the percentages of each contribution type across frames.

We observe that the distribution of types varies across frames. We see a significantly higher percentage of not-classified subjects in the Maintenance problems and a significantly higher percentage of conditional co-operators in the Provision problems. This is consistent with the findings of Gächter et al (2017) and Isler et al

[24] Note that there is a tendency to a greater disparity between average moral judgments of Provision and Maintenance problems when the other group member contributes extreme levels. For instance, a contribution of 10 when the other group member free rides is more praiseworthy in the Maintenance treatments. Also, a contribution of 30 when the other group member contributes 20 is seen as more praiseworthy on Maintenance treatments.

(2021), who find more conditional co-operators in Provision than Maintenance problems.

**Table 3.1.** *Percentage of each contribution type - per frame*

| Contribution Type | Provision | Maintenance | P-value (test of proportions) | P-value ($\chi^2$ test) |
|---|---|---|---|---|
| Free riders | 6.8% | 9.0% | 0.2527 | 0.315 |
| Conditional Co-operators | 70.4% | 62.9% | 0.0468** | 0.054* |
| Unconditional Co-operators | 9.9% | 10.4% | 0.8335 | 0.822 |
| Triangle Contributors | 6.2% | 4.3% | 0.3066 | 0.311 |
| Not Classified | 6.8% | 13.3% | 0.0077*** | 0.007*** |

*Notes: * p<0.1; ** p<0.05; *** p<0.01.*

However, unlike previous studies we do not find evidence of a higher percentage of free riders in Maintenance problems. The high percentage of conditional co-operators and low percentage of free riders is also consistent with the findings of Kocher et al (2008) who document a higher percentage of conditional co-operators in the US than in other countries (i.e., they found around 80% of conditional cooperation in the USA compared to around 40% of conditional co-operators in Austria and Japan).

### 3.4.3. *Effective contributions and beliefs of Provision and Maintenance problems*

We did not find any statistically significant difference in average effective contributions across frames (Mann-Whitney test, $p$ = 0.3677), whereas average beliefs were statistically significantly different (Mann-Whitney test, $p$ = 0.0030)[25].

The next three subsections use statistical analyses to answer three questions. First, do the moral motivations of blame avoidance and praise seeking influence the effective contribution preferences elicited in the P-experiment? Second, do the moral motivations of blame avoidance and praise seeking influence the effective contributions elicited in the *C-experiment*? Third, do the moral motivations of blame avoidance and praise seeking influence the behaviour of Provision and Maintenance problems differently? We answer the third question by carrying out all the statistical

[25] This result is robust to the non-parametric test used. The equality of medians test also yields similar qualitative results for median effective contributions ($\chi^2$ = 0.0001; $p$ = 0.991, ties split across groups) and beliefs ($\chi^2$ = 4.7422; $p$ = 0.029, ties split across groups). The two-sample Kolmogorov-Smirnov test for equality of distributions also yields the same qualitative results for effective contributions ($D$ = 0.0836, $p$ = 0.229) and beliefs ($D$ = 0.1664, $p$ = 0.000).

analyses separately for the Provision and Maintenance problems. The discourse of the following subsections is as follows.

### 3.4.4. *Do BA and PS explain effective contribution preferences?*

We use regression analysis (OLS estimates) as our main statistical tool to document the relation between the effective contribution preferences elicited in the *P-experiment* and the BA and PS procedures. As a baseline model, we regress (separately for the Provision and Maintenance problems) the effective contributions of the *conditional choice* task on the effective contribution of the other group member. The regression includes the observations of all subjects, meaning four observations per subject given that the *conditional choice* task elicits the preferred effective contribution per each of the feasible effective contributions of the other group member. To control for the dependency of the observations at the subject level, we cluster the standard errors at the subject level. We use this technique as it generalizes the one used to measure conditional cooperation in the literature (see, for example, the statistical methods of Gächter, Kölle and Quercia, 2017). The slope coefficient of this regression is interpreted as a measure of *strong reciprocity*.

We expand the baseline model by including the predictions of the BA and PS procedures for each observation as regressors. These variables measure the extent to which moral motivations influence effective contribution preferences in both decision problems. Additionally, we expand the previous model by including several controls and restricting the sample. This allows us to see the robustness of the effect of BA and PS, because we can test whether any significance of the BA and PS regressors is due to either (a) an under specification of the regression model or (b) including confused subjects in the regressions.

Table 3.2 presents the regression outputs for the models we introduced above. The results reveal three main insights. First, effective contribution preferences are influenced positively by the effective contribution of the other group member. This influence is slightly stronger in the Provision treatments. This replicates the findings of Gächter et al (2017) and Isler et al (2021), showing that many people are strong reciprocators.

**Table 3.2.** *Regressing effective contribution preferences on blame avoidance and praise seeking: Regression output*

| | **Dependent variable:** Effective contribution preferences – effective contributions in the conditional choice task | | | | | |
|---|---|---|---|---|---|---|
| | **Provision treatments** | | | **Maintenance treatments** | | |
| | (1) | (2) | (3) | (1’) | (2’) | (3’) |
| Constant | 6.099*** | 4.639*** | 2.598** | 6.846*** | 5.462*** | 3.456** |
| | (0.545) | (0.944) | (1.270) | (0.684) | (0.902) | (1.558) |
| $Other_i$ | 0.579*** | 0.533*** | 0.672*** | 0.530*** | 0.487*** | 0.635*** |
| | (0.029) | (0.034) | (0.035) | (0.034) | (0.039) | (0.046) |
| Blame Avoidance | | 0.135***$^a$ | 0.119**$^a$ | | 0.120**$^a$ | 0.139*$^a$ |
| | | (0.040) | (0.049) | | (0.057) | (0.071) |
| Praise Seeking | | 0.034 | 0.043 | | 0.052$^b$ | 0.032$^b$ |
| | | (0.037) | (0.050) | | (0.042) | (0.065) |
| Controls | No | No | Yes | No | No | Yes |
| MFQ variables | No | No | Yes | No | No | Yes |
| MAC-Q variables | No | No | Yes | No | No | Yes |
| Clusters | 324 | 324 | 231 | 279 | 279 | 179 |
| Adjusted R$^2$ | 0.32 | 0.33 | 0.49 | 0.24 | 0.26 | 0.40 |
| *p* value of *F* test$^a$ | | 0.000$^c$ | 0.020$^c$ | | 0.006$^c$ | 0.093$^c$ |
| AIC | 9,488.830 | 9,465.231 | 6,536.524 | 8,420.819 | 8,367.123 | 5,324.833 |
| BIC | 9,499.164 | 9,485.899 | 6,555.839 | 8,430.854 | 8,387.179 | 5,343.150 |

*Notes:* OLS estimates. We cluster the standard errors at the individual level (displayed in parentheses) as each regression includes the four elicited (via the strategy method) preferred effective contributions of each subject. Models (1), (2) and (3) run regressions including only subjects from the Provision treatments, whereas models (1’), (2’) and (3’) run those same models including only subjects from the Maintenance treatments. Models (1) and (1’) run the following equation: $cc_i = \beta_0 + \beta_1 \times Other_i + u_i$, where $cc_i$ represents the effective contribution of subject $i$ (subscript $i$ refers to a given subject) and $Other_i \in \{0, 10, 20, 30\}$ represents the effective contribution of the other group member. Models (2) and (2’) run the following equation: $cc_i = \beta_0 + \beta_1 \times Other_i + \beta_2 \times BA_i + \beta_3 PS_i + u_i$, where $BA_i$ and $PS_i$ represent, respectively, the Blame Avoidance and Praise Seeking predictions of subject $i$. Models (3) and (3’) add sociodemographic variables and all the moral foundations of MFQ and MAC-Q to the specification of models (2) and (2’). Additionally, models (3) and (3’) exclude both non-US subjects, subjects that failed the attention checks of MFQ and MAC-Q and subjects classified as hump-shaded or others in the strategy method task. * p<0.1; ** p<0.05; *** p<0.01. [a] The Blame Avoidance variable is statistically significant at least at the 5% level in either fixed-effects, random-effects, tobit or ordered probit panel data models.[b] The Praise Seeking variable, in the Maintenance treatments, is statistically significant at least at the 5% level in either fixed-effects, random-effects, tobit or ordered probit panel data models. [c] Joint test of significance of Blame Avoidance and Praise seeking ($H_0: \beta_2 = \beta_3 = 0$; $H_1: \beta_2 = \beta_3 \neq 0$).

Second, we find that blame avoidance is another driver of people’s effective contribution preferences in the Provision problem. In model (2), the positive coefficient of BA is statistically significant at the 1% level and the joint test of significance of both BA and PS is rejected at the 1% level.

Additionally, both goodness-of-fit tests unambiguously favour the model including the BA and PS procedures. Model (3) shows that BA is still statistically significant at the 5% level. Furthermore, the statistical significance of BA remains even when we use different ways to model the data: fixed-effects, random-effects, tobit and ordered probit panel data models.

Third, blame avoidance and praise seeking influence people's effective contribution preferences in the Maintenance problem. Model (2') shows a positive significant coefficient of BA, and the joint test of significance of BA and PS is rejected at least at the 10% level in models (2') and (3'). Statistical significance of BA remains, and PS becomes statistically significant, with all the remaining model specifications we used to analyse our data: fixed-effects, random-effects, tobit and ordered probit panel data models.

Taken together, these results reveal that both strong reciprocity and moral motivations are drivers of people's contribution preferences. Strong reciprocity has a larger effect than moral motivations on effective contribution preferences. Whilst strong reciprocity and BA are significant regressors in both decision problems, PS is only important in the Maintenance problem.

One issue with the previous analysis is that it does not control for the potential endogeneity of the effective contribution of the other group member, BA and PS. More specifically, it is conceivable, given Figure 3.2 above, that the effective contribution of the other group member influences the moral judgments of our subjects. If the variation in moral judgments is such that the prediction of BA and/or PS changes, then changes in the effective contribution of others will not only influence our dependent variable but also BA and PS.

Hence, one may wonder whether the statistical significance of BA or PS is an artefact of this endogeneity, and its effect indirectly pertains to the effective contribution of others. If this is the case, then BA and PS would not explain our subject's effective contribution preferences.

To control for the aforementioned issue, we analyse each effective contribution decision in the *conditional choice* task separately. More specifically, we analyse for each subject whether BA or PS successfully predict each effective contribution decision he or she made in the *conditional choice* task. This allows us to control for the endogeneity issue because looking at the different decisions in the *conditional choice* task separately holds the effective contribution of the other group members constant.

Additionally, and as a baseline, we analyse whether perfect conditional cooperation (making the same effective contribution as the other group member) successfully predicts each of the decisions in the *conditional choice* task. This baseline is akin to what Sugden's (1984) principle of reciprocity would predict under certain circumstances[26]. Also, this baseline can be seen as a proxy of the influence of the effective contribution of others in each decision in the conditional choice task.

Following the previous analysis, we calculate (separately for each decision in the *conditional choice* task) the percentage of subjects for which perfect conditional cooperation successfully predicts each decision. We also compute the percentage of subjects for which either BA or PS successfully predict a given decision.

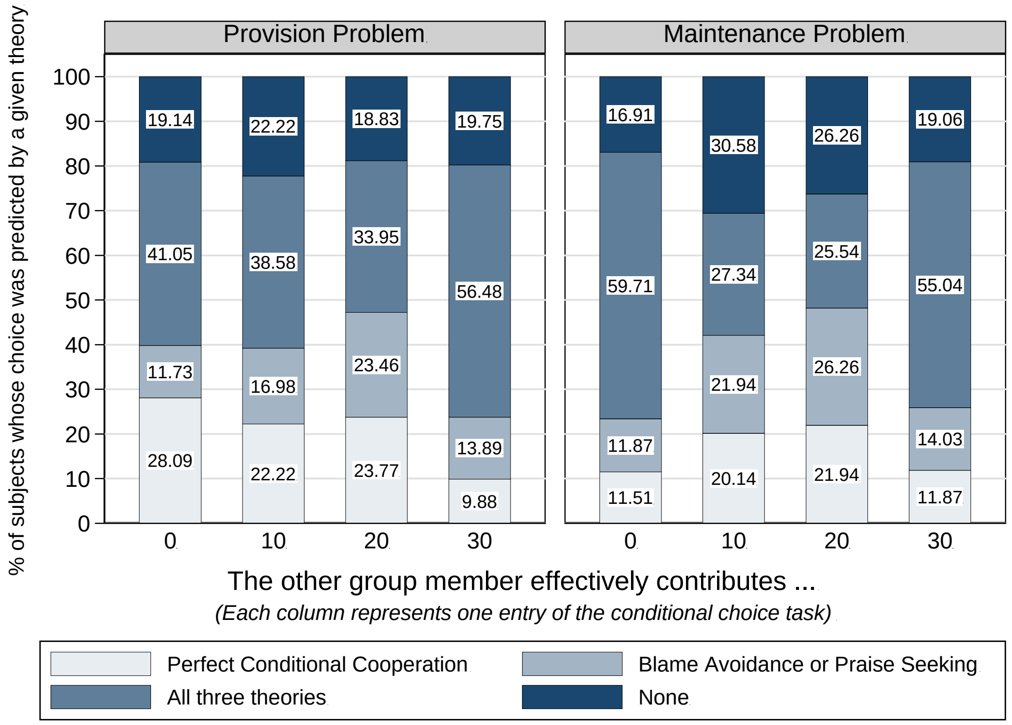


**Figure 3.3.** *Percentage of subjects' choices in each decision of the conditional choice task consistent with the analyse theories - per frame*

Figure 3.3 plots the percentage of subjects in one of four categories: (i) only consistent with perfect conditional cooperation; (ii) only consistent with either BA or PS; (iii) consistent with both perfect conditional cooperation and either BA or PS; (iv) inconsistent with the three procedures considered. Figure 3.3 plots, in different

[26] Perfect conditional cooperation would be predicted by Sugden (1984) if subjects' strategies are symmetric, and everyone would think that higher effective contribution is better than a lower effective contribution. Given the structure of the decision situations we analyse, the first condition holds. We take the moral judgments of Figure 3 as support for the second condition.

columns, the stacked percentages of these four categories for each decision in the *conditional choice* task. The left panel plots the statistical analysis for the Provision treatments and the right panel plots the statistical analysis for the Maintenance treatments.

Figure 3.3 presents our first main result. Moral motivations help to explain the decisions that subjects make in each of the decisions of the conditional choice task: between 11.7% and 23.5% of subjects' decisions in the Provision problem (and 11.9% to 26.2% in the Maintenance problem) are only consistent with at least one of the moral motivations we consider in this chapter. This reinforces our previous claims of the importance of moral motivations as drivers of effective contribution preferences. Even when we fix the effective contribution of others at a given level, and hence we control for the endogeneity issue mentioned earlier, moral motivations are the only theories to successfully predict the choices of a nontrivial percentage of subjects.

**Table 3.3.** *Percentage of subjects' choices in the conditional choice task predicted by each of the analysed theories*

| | % of subject's choices consistent with the analyzed theories | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | **Provision Treatments** | | | | **Maintenance treatments** | | | |
| | The other group member effectively contributes ($c_{-i}$) … | | | | | | | |
| | … 0 | … 10 | … 20 | … 30 | … 0 | … 10 | … 20 | … 30 |
| Only PCC | 28.1% | 22.2% | 23.8% | 9.9% | 11.5% | 20.1% | 21.9% | 11.9% |
| Only BA | 2.8% | 4.9% | 6.5% | 8.3% | 0.7% | 9.4% | 6.5% | 8.6% |
| Only PS | 8.3% | 10.2% | 15.1% | 1.9% | 9.7% | 7.9% | 15.1% | 1.8% |
| PCC and BA | 31.5% | 34.6% | 27.2% | 0.0% | 40.3% | 20.1% | 15.8% | 0.0% |
| PCC and PS | 0.0% | 0.3% | 3.1% | 33.0% | 0.0% | 4.7% | 5.8% | 42.8% |
| BA and PS | 0.6% | 1.9% | 1.9% | 3.7% | 1.4% | 4.7% | 4.7% | 3.6% |
| PCC and BA and PS | 9.6% | 3.7% | 3.7% | 23.5% | 19.4% | 2.5% | 4.0% | 12.2% |
| none | 19.1% | 22.2% | 18.8% | 19.8% | 16.9% | 30.6% | 26.3% | 19.1% |
| | Hit Rates of the three different theories | | | | | | | |
| PCC | 69.1% | 60.8% | 57.7% | 66.4% | 71.2% | 47.5% | 47.5% | 66.9% |
| BA | 44.4% | 45.1% | 39.2% | 35.5% | 61.9% | 36.7% | 30.9% | 24.5% |
| PS | 18.5% | 16.0% | 23.8% | 62.0% | 30.6% | 19.8% | 29.5% | 60.4% |
| None | 19.1% | 22.2% | 18.8% | 19.8% | 16.9% | 30.6% | 26.3% | 19.1% |

*Notes:* PCC = Perfect Conditional Cooperation; BA = Blame Avoidance; PS = Praise Seeking.

If we further disaggregate the data and consider BA and PS separately, as we do in Table 3.3, we observe four main features. First, Perfect Conditional Cooperation has a higher overall hit rate than either BA or PS. Second, the hit rates of Perfect Conditional Cooperation and BA are consistently higher than that of randomness (higher than 25%) in almost all the decisions of the conditional choice task. Third, the hit rate of BA is greater than that of PS for effective contributions of 0, 10 or 20, the reverse being correct for effective contributions of 30. Finally, for effective contributions of 0, 10 and 20, PS is the only theory that can predict the choices of 8.3% to 15.1% of the subjects in the Provision treatments (7.9% to 15.1% of the subjects in the Maintenance treatments). In other words, even when PS has a lower hit rate, it still adds to our understanding of the contribution preferences of a nontrivial minority of subjects. These four findings corroborate the qualitative findings of the regression models of Table 3.2.

### 3.4.5. *Do beliefs, ABC, BA, and PS explain effective contributions?*

The previous subsection documented a relation between effective contribution preferences and the BA and PS procedures. Given that ABC is constructed using effective contribution preferences, and that those are influenced by BA and PS, one may wonder to which extent ABC is capturing motives different from BA and PS. To answer this question, Figure 3.4 plots proportional Venn diagrams (one per each decision problem. Top panel, Provision problem; bottom panel, Maintenance problem)[27]. Each of the three ellipses represent the number of subjects for which a given procedure makes a correct prediction of their effective contributions in the *C-experiment.*

Each zone in the Venn diagrams is proportional to the number of subjects that fall within each zone, and, for reference, includes the percentage of the total subjects that lie inside it.

The Venn diagrams provide us with four insights. First, ABC has a substantially higher hit rate than either BA or PS (71.6% vs 34.6% vs. 38.6% in the Provision problem; 65.5% vs. 46.8% vs. 40.3% in the Maintenance problem). More importantly, the hit rates of all the theories are higher than what would be successfully predicted by randomness (25%)[28].

[27] We use the EulerAPE program presented in Micallef and Rodgers (2014) to generate our proportional Venn diagrams. To make each area of the Venn-Diagram proportional to the number of subjects that lie within it, the Venn-Diagrams had to use ellipses.

[28] As there exist four effective contribution levels in the unconditional choice task of the *C-experiment*, a predictive device that would give an equal chance to each effective contribution choice to be selected as the prediction (a uniform distribution randomness model) would have a 25% chance of getting an effective contribution choice right.

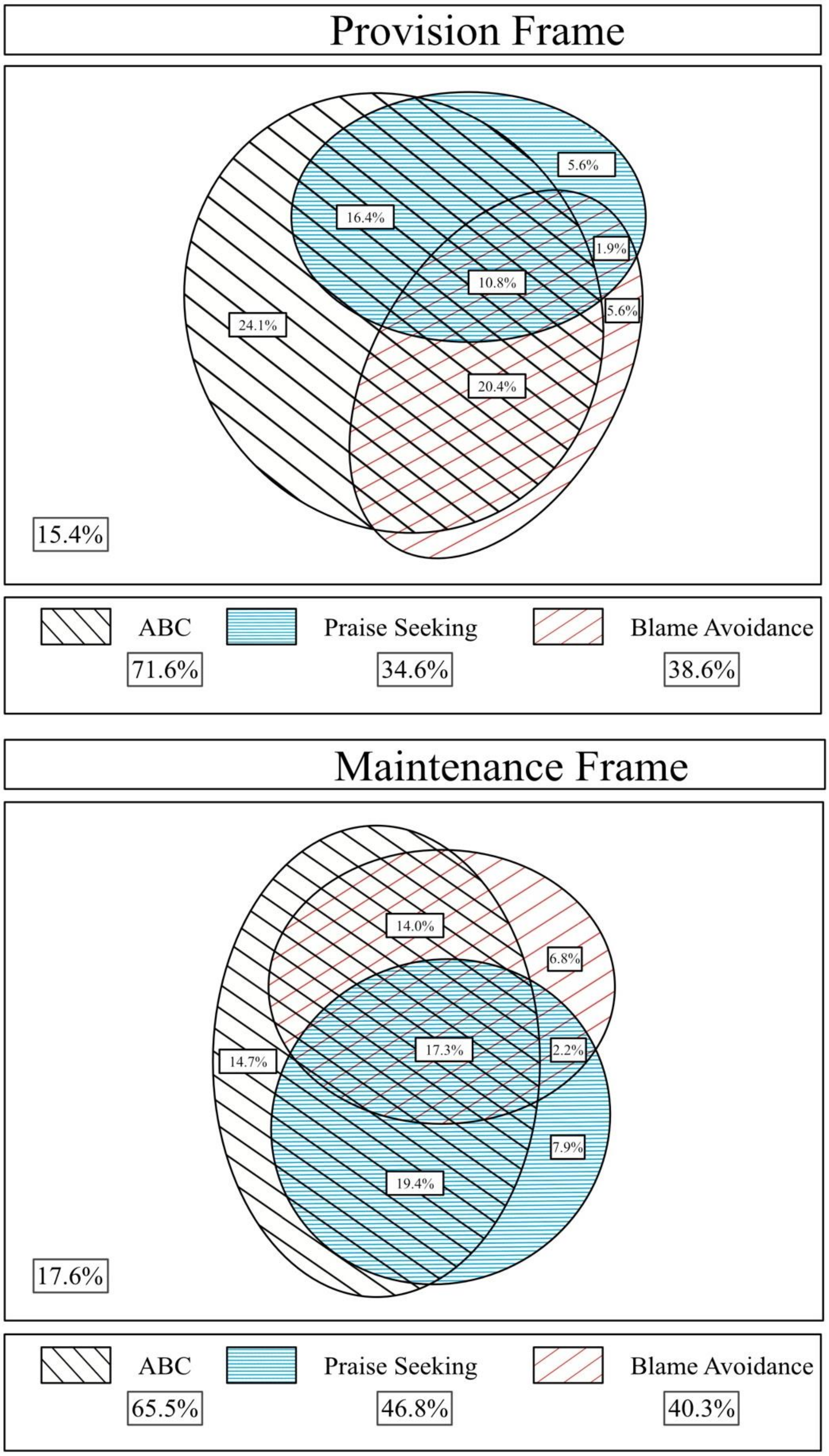


**Figure 3.4.** *Proportional Venn Diagrams (top panel - Provision frame, bottom panel - Maintenance frame) of the predictive success of the ABC, BA, and PS procedures*

Second, looking at the non-intersection zones of the ellipses we see that all the three procedures have instances of being the sole predictor of effective contribution choices. ABC has the higher hit rate of non-intersection zones (24.1% vs 5.6% vs. 5.6% in the Provision problem: 14.7% vs. 6.8% vs. 7.9% in the Maintenance problem). Third, both the overall and the non-intersection zones hit rates of ABC decrease in the Maintenance frame whilst those of the two moral motivations increase. This shows that moral motivations have a higher relative importance, when compared to ABC, in explaining effective contribution choices in the Maintenance problem.

Finally, and most importantly for the question raised before, a substantial number of subjects lie within the intersection zones between the three procedures (49.5% of subjects in the Provision frame; 52.9% of subjects in the Maintenance frame), especially in the intersection zones between the moral motivations and ABC.

This corroborates the claim that effective contribution preferences are influenced by BA and PS; and that, consequently, the predictions of ABC and those of either BA or PS are not mutually exclusive.

As a first approach to explaining effective contributions in the *C-experiment*, the remainder of the section will look at the importance of ABC, BA, and PS in driving effective contributions without considering the relation between ABC and the BA and PS procedures (which will be dealt with in the next subsection).

We use an econometric approach to measure the importance of ABC, BA, and PS in shaping our subjects' effective contributions in Provision and Maintenance problems. As a baseline model, and adapting Fischbacher and Gächter's (2010) regression approach, we regress the effective contributions in the *C-experiment* on the prediction from the ABC method and the beliefs about the effective contribution of the other group member. We expand this baseline model by including the predictions from the BA and PS procedures.

Our approach allows us to replicate previous results by using the baseline model, and to explore the extent to which the moral motivations captured by the BA and PS procedures influence effective contribution choices in the *C-experiment*. Finally, as a robustness check we include the moral foundations of MFQ and MAC-Q and several sociodemographic variables as controls in the previous model, and we further restrict the sample. This allows us to test whether BA and PS procedures predict the data because the moral motivations capture specific cognitive processes behind subjects' choices or simply because they are correlated with subjects' general moral worldviews, that can be correlated with behaviour but that have no underlying structure on how to map those views into predicted actions. We report the regression outputs in Table 3.4.

**Table 3.4.** *Regressing effective contributions in the C-experiment on Blame Avoidance and Praise Seeking: Regression Output*

| | **Dependent variable:** Effective contributions in the C-Experiment | | | | | |
|---|---|---|---|---|---|---|
| | **Provision treatments** | | | **Maintenance treatments** | | |
| | (1) | (2) | (3) | (1') | (2') | (3') |
| Constant | 4.078*** | 2.572** | 3.825 | 4.768*** | 2.892*** | 13.397** |
| | (0.758) | (0.995) | (4.039) | (0.936) | (0.962) | (6.473) |
| Belief | 0.244*** | 0.239*** | 0.020 | 0.363*** | 0.200** | 0.064 |
| | (0.079) | (0.081) | (0.089) | (0.076) | (0.080) | (0.127) |
| ABC | 0.563*** | 0.544*** | 0.747*** | 0.419*** | 0.398*** | 0.608*** |
| | (0.073) | (0.077) | (0.085) | (0.073) | (0.071) | (0.117) |
| Blame Avoidance | | 0.054 | 0.035 | | 0.220*** | 0.139* |
| | | (0.044) | (0.051) | | (0.066) | (0.074) |
| Praise Seeking | | 0.055 | 0.153** | | 0.116** | 0.144* |
| | | (0.040) | (0.061) | | (0.052) | (0.082) |
| Controls | No | No | Yes | No | No | Yes |
| MFQ variables | No | No | Yes | No | No | Yes |
| MAC-Q variables | No | No | Yes | No | No | Yes |
| Observations | 324 | 324 | 229 | 279 | 279 | 179 |
| Adjusted $R^2$ | 0.60 | 0.60 | 0.66 | 0.51 | 0.56 | 0.61 |
| *p* value of *F* test[a] | | 0.078 | 0.018 | | 0.000 | 0.019 |
| AIC | 2,153.616 | 2,151.868 | 1,530.274 | 1,972.622 | 1,947.135 | 1,260.783 |
| BIC | 2,164.958 | 2,170.772 | 1,633.286 | 1,983.515 | 1,965.291 | 1,356.405 |

*Notes:* OLS estimates. Robust standard errors reported in parentheses. Models (1), (2) and (3) run regressions including only subjects from the Provision treatments, whereas models (1'), (2') and (3') run those same models including only subjects from the Maintenance treatments. Models (1) and (1') run the following equation: $c_i = \beta_0 + \beta_1 \times belief_i + \beta_2 \times ABC_i + u_i$, where $c_i$ represents the effective contribution of subject $i$ in the C-experiment, $belief_i \in \{0, 10, 20, 30\}$ represents the belief of subject $i$ about the effective contribution level of the other group member and $ABC_i$ represents the prediction that the ABC method makes given the first-order belief and effective contribution preferences of subject $i$. Models (2) and (2') add, to the specification of models (1) and (1'), terms for the predictions of Blame Avoidance and Praise Seeking. Models (3) and (3') add sociodemographic variables and all the moral foundations of MFQ and MAC-Q to the specification of models (2) and (2'). Additionally, models (3) and (3') exclude both non-US subjects, subjects that failed the attention checks of MFQ and MAC-Q and subjects classified as hump-shaded or others in the strategy method task. * p<0.1; ** p<0.05; *** p<0.01. [a] Joint test of significance of Blame Avoidance and Praise seeking.

We have three main findings from the econometric models reported in Table 3.4. First, and in line with Fischbacher and Gächter (2010), the slope coefficient of the predicted effective contribution made by ABC in the baseline model is positive and significant, outlining the importance of ABC as an explanation of effective contribution choices in the *C-experiment*. A 1 token increase in the predicted effective contribution by ABC raises effective contributions, on average, by 0.56 (0.42) tokens in the Provision (Maintenance) problems. The significance and sign of the slope coefficient of ABC is robust to the inclusion of the BA and PS procedures and all the control variables. Also, the belief coefficient is positive and significant in both Provision and Maintenance treatments: a 1 token increase in beliefs of the other group member's contribution increases effective contributions by 0.24 (0.36) tokens in the Provision Maintenance treatments, and its significance and sign are robust to the inclusion of the BA and PS procedures. This shows that an extra reciprocity, on top of the conditional cooperation attitudes captured by the ABC approach, influences our subjects' behaviour in both frames.

Second, we find weak evidence favouring moral motivations as direct drivers of effective contribution choices in the Provision problem. Even when neither BA nor PS have a significant coefficient in model (2), the joint test of significance of both procedures is rejected at the 10% significance level. Furthermore, once we include all the controls and restrict our sample (see the note under table 4 for the specifics of the restriction rules we apply) PS becomes significant at the 5% level and the joint test of significance of both ethical theories is rejected at the 5% significance level as well. The slope coefficients of BA and PS, however, play a minor role in shaping effective contributions: their size is around 9% of the slope coefficient of ABC.

Third, we find strong evidence supporting both moral motivations as direct drivers of effective contribution choices in the Maintenance problem This claim is supported by the positive and significant coefficients of BA and PS in model (2'), by a rejection of the joint significance test of BA and PS at the 1% significance level in model (2') and by the survival of their statistical significance to the inclusion of several controls in model (3'). Furthermore, the relative size of the BA and PS slope coefficients with respect to ABC increases substantially when compared to Provision problems: the size of the slope coefficient of BA (PS) is around 55% (29%) the size of the slope coefficient of ABC.

Overall, this subsection provides evidence to support the view that the three procedures we consider motivate subjects' effective contribution choices. ABC plays a prominent role in shaping effective contributions, especially in the Provision problem. The BA and PS procedures have a substantial importance in directly shaping effective contributions in Maintenance problems.

### 3.4.6. *What is the total effect of BA and PS on effective contributions?*

As stated previously, the ABC approach and the moral motivations captured by BA and PS are not mutually exclusive. As BA and PS influence effective contribution preferences, which are at the core of the ABC method, this calls for a reformulation of the econometric model to account for such a relation. Otherwise, the coefficient of ABC in the regressions of Table 3.4 will be partially capturing an effect that is implicitly driven by BA and PS. Hence, the previous results are an accurate picture of the total effect of ABC on effective contributions. However, the picture of the effect of BA and PS in effective contributions is incomplete as the models in Table 4 do not reveal the indirect influence of BA and PS on effective contributions through effective contribution preferences. As such, the previous results should be interpreted with caution. As our interest lies on understanding the total effect that each of the procedures has in shaping subjects' effective contribution decisions, we use a mediation analysis to unravel the indirect effect of BA and PS in effective contribution choices.

The mediation analysis we use estimates two equations. The first equation is model (2) in Table 3.4 (for the Provision frame, and (2') for the Maintenance frame): we regress the effective contributions in the *C-experiment* on beliefs and the effective contributions predicted by the ABC, BA, and PS procedures. The second equation regresses the effective contribution predicted by ABC on the effective contributions predicted by BA and PS. There are several reasons for the choice of this second equation. First, the ABC procedure is already constructed by using the belief of the effective contribution of the other group member. Hence, it is redundant to include beliefs in the equation. Any extra effect of beliefs on the effective contributions in the *C-experiment* should uniquely be captured by its relevant coefficient in the first equation, which can be interpreted as an extra conditional cooperation attitude on top of the one coming from effective contribution preferences. Second, the ABC method is based on the usage of effective contribution preferences, which – as we know from subsection 3.4.4. – are influenced by BA and PS. So, *a priori* we would expect there to be a relation between the predictions of BA and PS and the predictions of ABC. We take such a relation as a proxy for the underlying relation between effective contribution preferences and the moral motivations captured by BA and PS. To summarise, the mediation analysis simultaneously estimates the following two structural equations using the maximum likelihood method:

$$C_i = \beta_0 + \beta_1 \times b_i + \beta_2 \times ABC_i + \beta_3 \times BA_i + \beta_4 \times PS_i + u_i \tag{3.3}$$

$$ABC_i = \delta_0 + \delta_1 \times BA_i + \delta_2 \times PS_i + \varepsilon_i \tag{3.4}$$

Substituting the second equation into the first one and doing some simple algebraic rearrangement, we can rewrite the first equation in terms of the total effect of blame avoidance and praise seeking on the effective contributions in the *C-experiment*:

$$C_i = \beta_0 + \beta_1 \times b_i + \beta_2 \times (\delta_0 + \varepsilon_i) + (\beta_3 + \beta_2 \times \delta_1) \times BA_i + (\beta_4 + \beta_2 \times \delta_2) \times PS_i + u_i \tag{3.5}$$

where $\beta_2$ is the total effect of ABC on effective contributions; $\beta_3$ and $\beta_4$ are the direct effects of BA and PS on effective contributions; $\beta_2 \times \delta_1$ and $\beta_2 \times \delta_2$ are the indirect effects of BA and PS on effective contributions through ABC and $(\beta_3 + \beta_2 \times \delta_1)$ and $(\beta_4 + \beta_2 \times \delta_2)$ are the total effects of BA and PS on effective contributions.

Figure 3.5 plots the results of the mediation analysis for the Provision problem and Figure 3.6 for the Maintenance problem. The top panels of Figures 3.5 and 3.6 present the estimates for the six relevant coefficients ($\beta_1$, $\beta_2$, $\beta_3$, $\beta_4$, $\delta_1$ and $\delta_2$) in a path diagram and the bottom panels show the total effects of each of the four variables (beliefs, ABC, blame avoidance and praise seeking) disaggregated into direct and indirect effects. Both beliefs and ABC only have a direct effect whereas both blame avoidance and praise seeking have indirect effects as well. The numbers above the bars represent the value of the total effect and the numbers within the bars represent the value of the indirect effect. The grey-coloured parts of the bar representing the total effect of ABC correspond to the indirect effect of either BA or PS (light shading: blame avoidance; dark shading: praise seeking). We plot the indirect effects of BA and PS with the same shades of grey used in the *total-effects-bar* of ABC to stress the connection between them. The grey-coloured parts of the bars are to be interpreted as mediated effects: the effect of blame avoidance and praise seeking that is mediated by ABC. We include stars signifying statistical significance (* 10%, ** 5%, *** 1%

significance levels) in both the top and the bottom panels. The mediation analysis modifies the results of the previous section in several respects.

First, the mediation analysis reveals a stronger effect than the one reported in the previous section of BA and PS on the effective contributions in the *C-experiment*. Regardless of the framing of the decision situation, there is at least one indirect effect that is statistically significant.

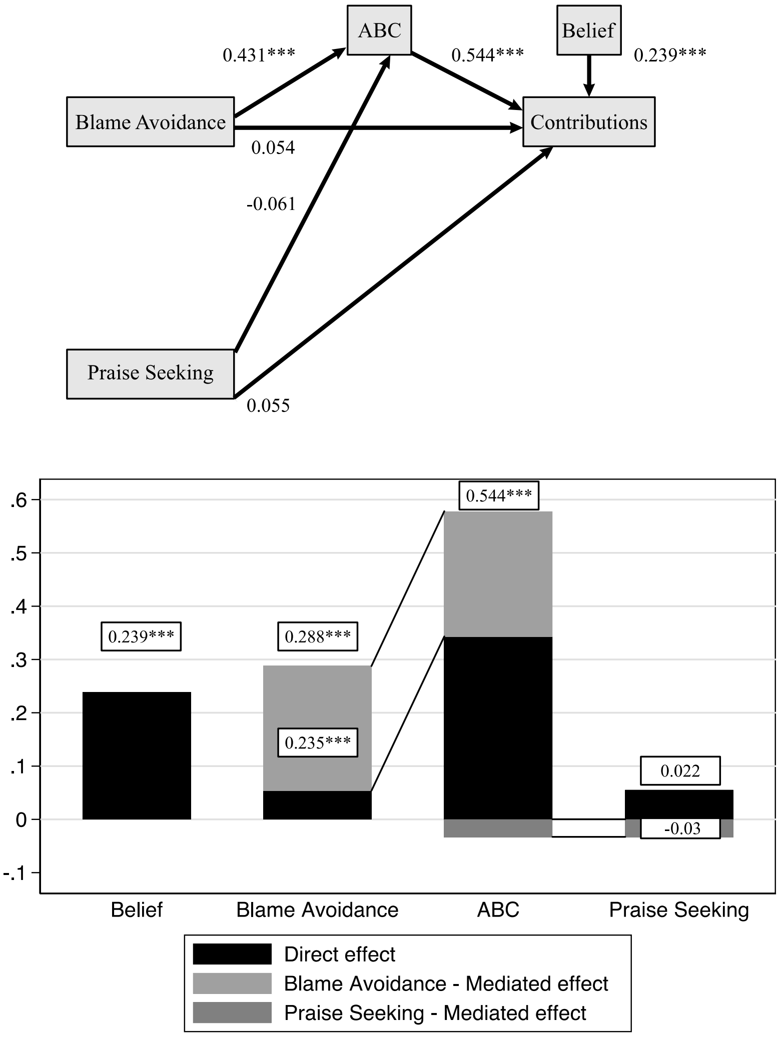


**Figure 3.5.** *Regression coefficients (top panel) and disaggregated total effects (bottom panel) of the mediation analysis - Provision treatments*

Second, and as can be seen in Figure 3.5, BA proves to be an important motive behind the effective contribution choices in the Provision frame. Moreover, its importance relies solely on the indirect effect that it has on effective contributions: BA has a statistically significant indirect effect, representing 43.20% ($\frac{0.235}{0.544} = \delta_1 = 0.432$

of the total effect of ABC on effective contributions. This is a huge increase compared to the previous section, where the slope coefficient of BA represented around 9% of the slope coefficient of ABC.

This effect dramatically changes the picture depicted by the previous section. It is not the case that moral motivations do not drive effective contributions in the *C-experiment* of the Provision problem; but, rather, that they do so by influencing the subjects' effective contribution preferences. This is qualitatively consistent with the findings of subsection 3.4.4., where model (2) of Table 3.2 showed a positive relation between BA and effective contribution preferences.

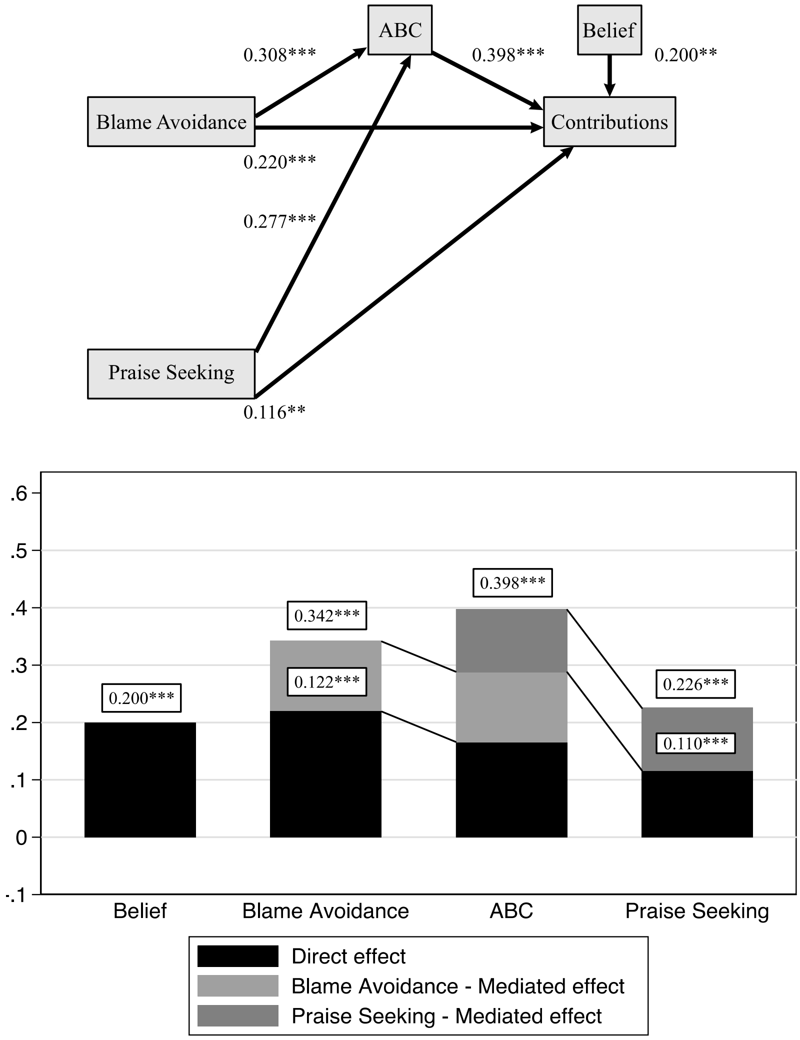


**Figure 3.6.** *Regression coefficients (top panel) and disaggregated total effects (bottom panel) of the mediation analysis - Maintenance treatments*

Third, results presented in Figure 3.6 point out that both BA and PS are important drivers of effective contribution choices in the Maintenance frame. Unlike in the Provision frame, both direct effects of BA and PS are statistically significant in the Maintenance frame. On top of them, both indirect effects are statistically significant as well. 30.65% ($\frac{0.122}{0.398} \approx 0.3065 \delta_1$)[29] of the total effect of ABC captures an underlying effect of BA on effective contribution preferences, whereas 27.64% ($\frac{0.110}{0.398} = \delta_2 \approx 0.2764$) of the total effect of ABC captures an underlying effect of PS on effective contribution preferences.

Added together, 58.24% of the total effect that ABC has on effective contributions is capturing the underlying effects that BA and PS have on effective contribution preferences. Again, these results are consistent with the discussion in subsection 3.4.4.

Finally, the mediation analyses show that people rely more on their moral motivations, as captured by BA and PS, when making their effective contribution choices in the Maintenance frame. Whereas in the Provision frame PS's total effect on effective contributions is practically non-existent, in the Maintenance frame its total effect is more than half of the total effect of ABC ($\frac{0.226}{0.398} \times 100 = 56.78\%$). Additionally, the total effect of BA increases from 0.288 in the Provision problem to 0.342 in the Maintenance problem (cf. lower panels of Figs. 3.5 and 3.6).

Overall, the mediation analysis confirms that ABC, BA, and PS are important motives that drive people's effective contribution choices. It also allows us to identify the cause of the big effect size of ABC: ABC mediates the effect of the moral motivations captured in BA and PS. Once we account for that mediation, both BA and PS are substantial drivers of behaviour in public goods games. Without them, we fail to understand not only a significant proportion of subjects' choices but also the reasons as to why effective contribution preferences make ABC so successful in predicting effective contributions in the *C-experiment*.

[29] Note that, if we divide the indirect effect of blame avoidance by the direct effect of ABC, we get the coefficient $\delta_1$: $\frac{\beta_2 \times \delta_1}{\beta_2} = \delta_1$. Hence, $\delta_1$ can also be interpreted accordingly.

## 3.5. Concluding remarks and discussion

In this chapter, we investigate (i) the role of the effective contribution of the other group member and the moral motivations of blame avoidance and praise seeking, as captured by BA and PS respectively, in shaping effective contribution preferences; (ii) the role of beliefs, effective contribution preferences and the moral motivations of blame avoidance and praise seeking in driving effective contributions. We do this for two different frames of a public good problem: the Provision and the Maintenance problems. To do that, we elicit effective contribution preferences with a *conditional choice* task in the *P-experiment*; we elicit beliefs with an incentivised belief task in the *C-experiment*; and we measure the moral perception of our subjects, from an impartial spectator viewpoint, for all the strategy combinations of the decision problem he or she is confronted with (*M-experiment*). We then use this data to generate predictions using three procedures: ABC, BA and PS. The ABC approach captures the potential effects of effective contribution preferences and beliefs in effective contributions and BA and PS procedures capture the potential effects of the moral motivations of blame avoidance and praise seeking in effective contributions. We use the predictions of ABC and the predictions of BA and PS to measure their importance in shaping effective contribution decisions (*C-experiment*). Additionally, we use the predictions of BA and PS to measure the importance of moral motivations in shaping effective contribution preferences (*P-experiment*).

We find that moral motivations are an important driver of effective contribution preferences. On top of the effect of the effective contribution of the other group member, we observe that blame avoidance drives people's effective contribution preferences in both the Provision and Maintenance problems. Additionally, we observe that praise seeking potentially drives effective contribution preferences in the Maintenance problem. Additionally, we find that beliefs, effective contribution preferences, blame avoidance and praise seeking influence people's effective contributions in the *C-experiment*. More specifically, beliefs, effective contribution preferences and blame avoidance are important motives in shaping effective contributions in both the Provision and Maintenance problems, and praise seeking is mostly important in Maintenance problems. Moreover, we find that a great deal of the importance of effective contribution preferences, and hence the ABC method, in shaping effective contributions is due to the influence that moral motivations have in shaping those effective contribution preferences in the first place. In other words, moral motivations play a dual role: that of shaping effective contributions directly and that of shaping effective contribution preferences, which are a determinant of effective contributions.

In both economics and philosophy, there has been a debate as to whether all prosocial actions are fully reducible to personal considerations or not. Amartya Sen (1977) distinguishes between sympathy and commitment, the former including the welfare of others within one's own welfare as an explanation of prosociality and the latter driving prosocial actions without any consideration to one's own welfare, however broadly construed. Sen (1977) comments that (i) 'Commitment is, of course, closely connected with one's morals' (pp.329); (ii) 'one area in which the question of commitment is most important is that of the so-called public goods' (pp.330); (iii) 'commitment does involve, in a very real sense, counterpreferential choice' (pp. 328). Almost all social preferences models in economics are encapsulated in choices driven by preferences and, as such, they keep the *Sympathy* paradigm as the driving mechanism explaining people's actions.

Even the most recent models including moral concerns as a driving force of behaviour are based on preferential choice (e.g. Alger and Weibull, 2013). However, some authors advocate for a counterpreferential approach to prosociality. Smith and Wilson (2017, 2019) present a model where moral judgments of praiseworthiness and blameworthiness of an action change the probability of choosing such action, rather than changing one's own utility attached to that action. They propose blame and praise as two different mechanisms driving people's behaviour. Recognising Sen's (1977) and Smith and Wilson's (2019) arguments supporting the link between morality and behaviour we build two procedures, BA, and PS, to investigate the role of the moral motivations of blame avoidance and praise seeking in shaping people's contribution choices in public goods games.

Regardless of whether the moral motivations we consider are utility-based or channelled through non-preference paths, such as rules or principles governing behaviour, what is clear is that they do influence people's effective contributions and effective contribution preferences in public good games. Our analysis leaves as an open question how it is best to model moral motivations for decision making. However, some features of our data (e.g., the role of blame avoidance in shaping effective contribution decisions) suggest economists may find it useful to investigate the relevance of non-utility maximisation models to explain prosocial behaviour.

# Chapter 4. Moral Rules and Social Preferences in Cooperation problems

## 4.1. Introduction

The objective of my chapter is to study whether, and if so how, moral judgments and social preferences influence cooperation 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 cooperation 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 cooperation 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 cooperation 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 cooperation attitudes in both public goods problems. By eliciting for each experimental subject the cooperation attitudes in both cooperation 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 cooperation 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)[30]. 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)[31]. More recent research shows that people's cooperation attitudes 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, Fischbacher and Gächter, 2010, and Thöni and Volk, 2018)[32].

Despite the wide range of social preferences that can explain cooperation 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 cooperation attitudes are scarce, let alone i) tests that compare several theories at the same time, and ii) tests that analyse a theory's predictive success at the individual level

[30] See Bohm (1972); Dawes et al (1977); Marwell and Ames (1979) and Isaac et al (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.

[31] For empirical evidence, see, as well, Andreoni (1988, 1990 and 1995); Croson (1996); Ferraro and Vossler (2010); Palfrey and Prisbey (1996 and 1997); Anderson et al (1998). For theoretical models built to accommodate this evidence, 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.

[32] For literature on cooperation 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.

(but see, for instance, Beranek et al, 2017 for a within-subjects test of inequality aversion's predictive power of cooperation 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 cooperation 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 cooperation attitudes at the individual level.

Additionally, by examining jointly the cooperation attitudes in social dilemmas and common interest games allows a theoretical separation between the predictions of the considered theories[33]. 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 cooperation attitudes that we conjectured, ex ante, to be prevalent among subjects (conditional cooperation in the social dilemma and unconditional cooperation in the common 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 cooperation attitudes. Besides the theoretical usefulness of studying common interest games, they represent real life cooperation 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 chapter 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[34]. Morality has been

[33] This is highlighted in Palfrey and Prisbey (1997, see especially the discussion in pp. 830-831). But switching the focus to cooperation attitudes makes the theoretical separation more interesting, as it allows me to differentiate between different social preference models (see section 4 and Appendix C.2).

[34] Works that have influenced my view on the topic and prosociality and driven me to study morality are those of Aristotle (2000); Thomas Hobbes (1996 and 2008); the Earl of Shaftesbury (2000); Francis Hutcheson (2002 and 2004); David Hume (1739 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 (2008) for good discussions on the relation between

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, 1739 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.

The statistical analysis of the experimental games indicates that cooperation attitudes 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 cooperation 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 cooperation attitudes in both games. More

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); Nucci (1996); De Waal (1997); 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.

specifically, blame avoidance, maximin, and inequality aversion motives are the major determinants of cooperation attitudes in social dilemmas and common interest games. Reciprocity, social efficiency, praise seeking, and material selfishness are only determinants of cooperation attitudes in common interest games, and spite is not a determinant of cooperation attitudes in either cooperation problem.

The chapter proceeds as follows. Section 4.2 presents the experimental design. Section 4.3 presents the novel theoretical framework and its theoretical predictions. Section 4.4 discusses the theoretical predictions of the social preference models I consider. Section 4.5 presents the results of my experiment and section 4.6 concludes.

## 4.2. 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 *cooperation 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 chapter, 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.

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 cooperation 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 analysing cooperation attitudes. Third, by eliciting the P-experiments, M-experiments, and the parameters for each subject I was able to get each subject's observed cooperation attitudes of the SDG and the CIG and the predictions that each of the considered models make for those cooperation 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 chapter is to study the motivations behind cooperation 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.

### 4.2.1. *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:

$$30 - c_i + m * (c_i + c_{-i}) \tag{4.1}$$

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[35]. 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).

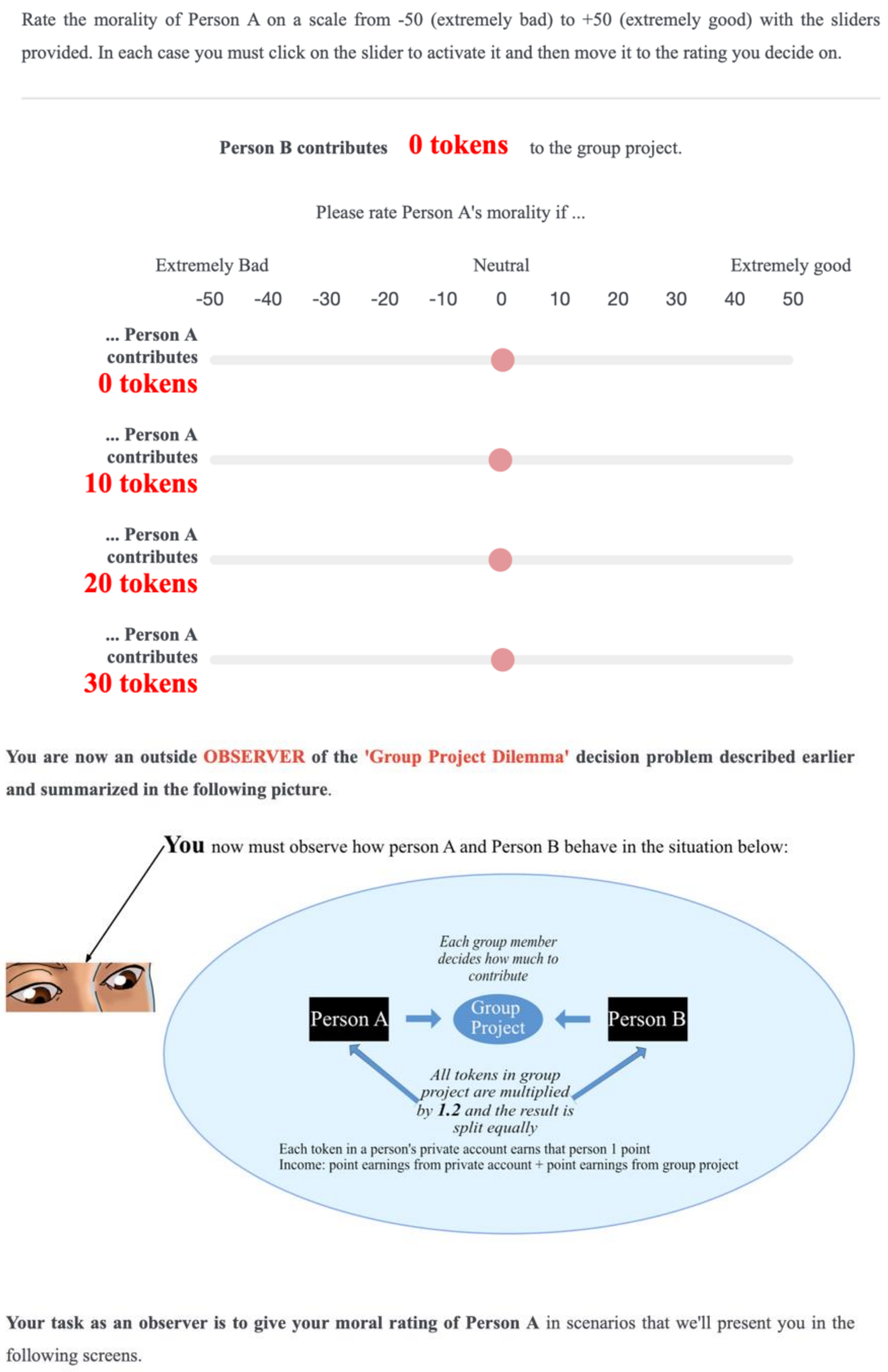
### 4.2.2. *Experimental tasks*

#### 4.2.2.1. *The M-experiments*

I use the survey method introduced by Cubitt et al (2011), and used in previous chapters 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 4.1.a 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.

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



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**

**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.

**Figure 4.1.** *(a) Top Panel: Screenshots of scenarios; (b) Bottom Panel: Implementation of the Impartial Spectator feature*

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 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 4.1.b summarizes how I introduced this feature to subjects in the M-experiment for the SDG.

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 chapter 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).

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)[36]. 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.

[36] 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).

#### 4.2.2.2. *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 cooperation 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)[37], 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 $c_i^* = c_{-i} \forall c_{-i}$; and an unconditional cooperator as a subject whose contributions are of the type $c_i^* = 30 \forall c_{-i}$.

#### 4.2.2.3. *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

[37] 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.

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 $x > P$[38].

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 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.

[38] 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).

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 Appendix C.2 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.

#### 4.2.2.4. *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.

### 4.2.3. *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 chapter (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.

## 4.3. The MRC framework: from Morality to Rules to Choices

### 4.3.1. *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 trade-offs (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).

### 4.3.2. *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 4.1.a below presents the normal form matrix of the social dilemma game under the assumption that both players' utility depends 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 4.1.b 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 cooperation in the social dilemma game; and, crucially, the prediction will depend on the strength of a subject's aversion towards advantageous inequality.

**Table 4.1.** *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**

| $i \setminus -i$ | $c_{-i} = 0$ | $c_{-i} = 10$ | $c_{-i} = 20$ | $c_{-i} = 30$ |
|---|---|---|---|---|
| $c_i = 0$ | 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**

| $i \setminus -i$ | $c_{-i} = 0$ | $c_{-i} = 10$ | $c_{-i} = 20$ | $c_{-i} = 30$ |
|---|---|---|---|---|
| $c_i = 0$ | 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 4.2, 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 4.2.** *i's Moral judgments of Person A in the SDG*

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

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

The first evident difference with classical models of social preferences is that the matrix in Table 4.2 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 4.3 from Table 4.2, where the moral judgments are kept the same, but now the players are $i$ and $-i$.

**Table 4.3.** *i's Moral judgments of him/herself in the SDG*

| $\boldsymbol{i}$**'s Moral judgments of** $\boldsymbol{i}$ **in the Social Dilemma Game …** | | | | |
|---|---|---|---|---|
| $i \backslash -i$ | $c_{-i} = 0$ | $c_{-i} = 10$ | $c_{-i} = 20$ | $c_{-i} = 30$ |
| $c_i = 0$ | $-3$ | $-15$ | $-25$ | $-34$ |
| $c_i = 10$ | $+12$ | $+7$ | $-8$ | $-17$ |
| $c_i = 20$ | $+24$ | $+20$ | $+12$ | $-2$ |
| $c_i = 30$ | $+37$ | $+32$ | $+29$ | $+20$ |

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 4.1.a matrix in Tables 4.4.a and 4.4.b.

**Table 4.4.** *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 |

Table 4.4.a represents the normal form matrix of the SDG with all the cells representing strategy combinations not suggested by blame avoidance shaded in black. Similarly, Table 4.4.b 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 black 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).

Whenever a moral rule suggests a single strategy to be taken, as is the case with praise seeking in Table 4.4.b, 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 4.4.a, 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 $c_{-i} = 30$.

### 4.3.3. *A formal presentation of the MRC framework: praise seeking and blame avoidance*

#### 4.3.3.1. *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 $M \coloneqq \{-50, \ldots., 0, \ldots., +50\}$ 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.

### 4.3.3.2. *Assumptions of the MRC framework and predictions*

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 cooperation 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 (1739, 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:

$$If\ \langle c_a, c_b\rangle = \langle c_i, c_{-i}\rangle, then\ m(\langle c_a, c_b\rangle, g, i) \equiv m_i(\langle c_i, c_{-i}\rangle, g) \quad (4.2)$$

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 4.3.2, where I go from Table 4.2 to Table 4.3). 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:

$$\frac{\partial m(\langle c_a, c_b \rangle, g, i)}{\partial i} \gtreqless 0 \tag{4.3}$$

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,

$$If\ m(\langle c_a, c_b \rangle, g, i) \neq m(\langle c_a, c_b \rangle, g, -i), then\ m_i(\langle c_i, c_{-i} \rangle, g) \neq m_{-i}(\langle c_i, c_{-i} \rangle, g) \tag{4.4}$$

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 (1739), 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*[39]. It is this assumption that makes the

[39] 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*"

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:

$$(4.5) \qquad B_{i,c_{-i},g} \coloneqq \left\{c_i \in C \mid (\forall c_i' \in C)\big(m_i(\langle c_i, c_{-i}\rangle, g) \geq m_i(\langle c_i', c_{-i}\rangle, g)\big)\right\}$$

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 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:

$$(4.6) \qquad U_{i,c_{-i},g} \coloneqq \{c_i \in C | m_i(\langle c_i, c_{-i}\rangle, g) \geq 0\},$$

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:

(4.7) $$c_i \succ c_i' \; iff \; \pi_i(\langle c_i, c_{-i}\rangle, g) > \pi_i(\langle c_i', c_{-i}\rangle, g)$$

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,

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

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

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:

(4.10) $$PS_{i,c_{-i},g} := \begin{cases} B_{i,c_{-i},g} \; if \; |B_{i,c_{-i},g}| = 1 \\ B'_{i,c_{-i},g} \; if \; |B_{i,c_{-i},g}| > 1 \end{cases}$$

(4.11) $$BA_{i,c_{-i},g} := \begin{cases} B'_{i,c_{-i},g} \; if \; U_{i,c_{-i},g} = \emptyset \\ U_{i,c_{-i},g} \; if \; |U_{i,c_{-i},g}| = 1 \\ U'_{i,c_{-i},g} \; if \; |U_{i,c_{-i},g}| > 1 \end{cases}$$

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:

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

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

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 cooperation attitudes in the SDG and CIG.

## 4.4. Social preferences and cooperation 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 cooperation attitudes. In this section I present the intuition behind the theoretical predictions of cooperation attitudes that the material selfishness, inequality aversion and sequential reciprocity models make, relegating the proofs to Appendix C.2. Additionally, I present the other social preference models I use, but relegate all the discussion on their theoretical predictions of cooperation attitudes to Appendix C.2.

### 4.4.1. *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)

### 4.4.2. *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:

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

Where $\pi_i$ and $\pi_{-i}$ denote the payoffs of subject $i$ and the other subject in the interaction, and the parameters $\alpha_i$ and $\beta_i$ 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 cooperation attitudes will be*

*(i), in the Social Dilemma,*

*(a) be a free rider ($c_i^* = 0\ \forall c_{-i} \in C$) iff $\beta_i < 1 - m$*

*(b) be a perfect conditional co-operator ($c_i^* = c_{-i}\ \forall c_{-i} \in C$) iff $\beta_i > 1 - m$*

*(c) be indifferent between $c_i \in [0, c_{-i}]$ iff $\beta_i = 1 - m$*

*(ii), in the Common Interest Game,*

*(a) be an unconditional co-operator ($c_i^* = 30\ \forall c_{-i} \in C$) iff $\alpha_i < m - 1$*

*(b) be a perfect conditional co-operator ($c_i^* = c_{-i}\ \forall c_{-i} \in C$) iff $\alpha_i > m - 1$*

*(c) Be indifferent between $c_i \in [c_{-i}, 30]$ iff $\alpha_i = m - 1$*

*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.

### 4.4.3. *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:

$$(4.15)\quad U_i^{\mathrm{DK}}(\pi_i, \pi_{-i}) \coloneqq \pi_i\left(a_i(h), b_{i,-i}(h)\right) + Y_{i,-i} * \kappa_{i,-i}\left(a_i(h), b_{i,-i}(h)\right) * \lambda_{i,-i,i}\left(b_{i,-i}(h), c_{i,-i,i}(h)\right)$$

Where $\pi_i$ 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* $U_i^{DK}(\pi_i(c_i, c_{-i}), \pi_{-i}(c_i, c_{-i}))$, *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* $Y_{i,-i} < \frac{1-m}{m^2 \times (c_{-i}-15)}$

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

*(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* $Y_{ij} > \frac{m-1}{m^2 \times (30-c_{-i})}$

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

*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 cooperation in the common interest game.

#### 4.4.4. *Other social preference models: spitefulness, social efficiency, and maximin*

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

$$U_i^{\mathrm{S}}(\pi_i, \pi_{-i}) \coloneqq \pi_i - \beta_i \times Max\{\pi_i - \pi_{-i}, 0\} \tag{4.16}$$

$$U_i^{\mathrm{SE}}(\pi_i, \pi_{-i}) \coloneqq (1 - p_i)\pi_i + p_i \times (\pi_i + \pi_{-i}) \tag{4.17}$$

$$U_i^{\mathrm{MM}}(\pi_i, \pi_{-i}) \coloneqq (1 - q_i)\pi_i + q_i \times Min\{\pi_i, \pi_{-i}\} \tag{4.18}$$

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

## 4.5. Results

### 4.5.1. *Descriptive statistics*

#### 4.5.1.1. *Average moral judgments of cooperation problems*

Figure 4.2 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 chapters, 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.

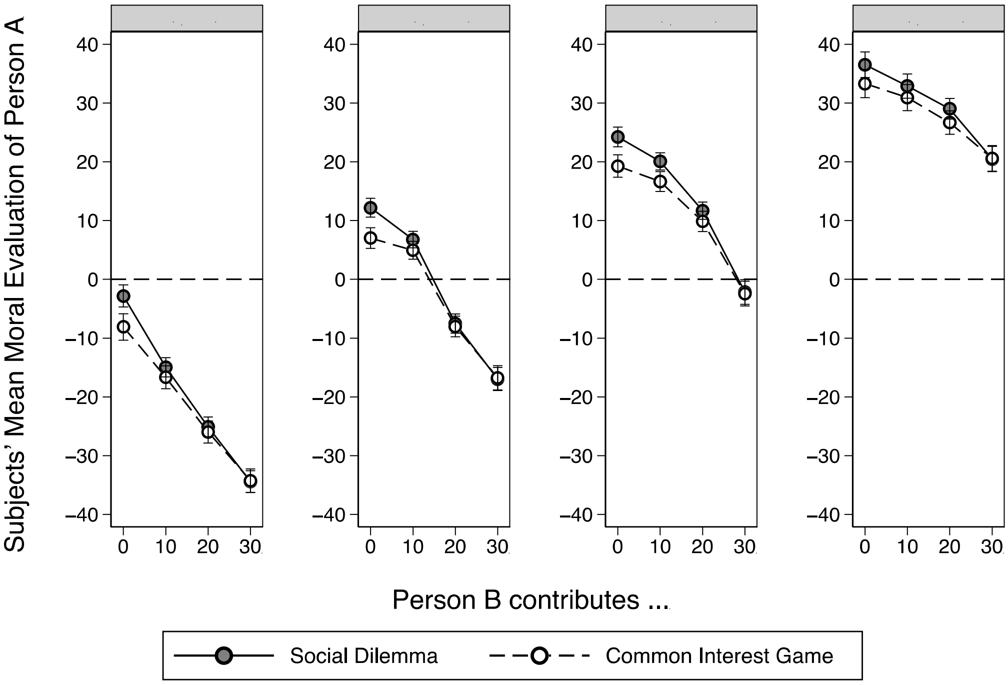


**Figure 4.2.** *Moral Evaluation Functions of all contributions of Person A*

Four features of Figure 4.2 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).

I now discuss what the average moral judgments in Figure 4.2 reveal about the predicted cooperation attitudes of praise seeking and blame avoidance in the SDG and CIG. Using the notation introduced in subsection 4.3, I can describe the predicted cooperation attitudes of praise seeking as:

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

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

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 cooperation. Regarding blame avoidance, I can describe its predicted cooperation attitudes in the $SDG$ and the $CIG$ as:

(4.21) $$\overrightarrow{BA}_{i,SDG} = \left(BA_{i,0,SDG}, BA_{i,10,SDG}, BA_{i,20,SDG}, BA_{i,30,SDG}\right) = (10,10,20,30)$$

(4.22) $$\overrightarrow{BA}_{i,CIG} = \left(BA_{i,0,CIG}, BA_{i,10,CIG}, BA_{i,20,CIG}, BA_{i,30,CIG}\right) = (30,30,30,30)$$

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 cooperation. 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.

#### 4.5.1.2. *Cooperation attitudes of cooperation problems*

I now report in Table 4.5 the distribution of subjects' cooperation 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 cooperation attitudes varies across games; and, incidentally, allows me to compare those observed cooperation attitudes with the predicted cooperation attitudes of praise seeking and blame avoidance given the moral judgments of Figure 4.2. I classify cooperation 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 4.5.** *Distribution of contribution types in cooperation problems*

| | Social dilemma | Common interest game | $\chi^2$ | 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 cooperation, which makes blame avoidance, before any analysis, a good candidate to predict cooperation attitudes in the SDG. In the common interest games, the two most common types are conditional cooperation (approx. 58%) and unconditional cooperation (approx. 34%). Nonparametric $\chi^2$ 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 cooperation to unconditional cooperation is predicted by the average cooperation attitudes of blame avoidance, proving it as a good candidate to fit the data. Praise seeking, by predicting unconditional cooperation in both games, is *ex ante* better suited to be a determinant of cooperation attitudes in the CIG.

I additionally report the joint distribution of types in Table 4.6. This analysis complements the previous one as it allows me to determine whether cooperation 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 cooperation 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 cooperation attitudes in the SDG and CIG, showing that for a substantial amount of the sample cooperation attitudes are specific to the cooperation problem.

**Table 4.6.** *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 | 0.63% | 5.97% | 4.09% | 0.00% | 0.31% |
| | Unconditional co-operators | 0.00% | 2.52% | 0.00% | 0.00% | 0.00% |
| | Conditional co-operators | 0.31% | 25.16% | 50.31% | 1.57% | 2.83% |
| | Hump shaded | 0.00% | 0.00% | 2.52% | 1.89% | 0.63% |
| | Other | 0.00% | 0.00% | 0.63% | 0.00% | 0.63% |

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 cooperation 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 cooperation 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 cooperation in the SDG could not be compatible with unconditional cooperation in the CIG for inequality aversion and reciprocity. Also, social efficiency and spite are not compatible with conditional cooperation 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.

### 4.5.1.3. *Parameter estimates of social preference models*

I end subsection 4.5 by presenting in Table 4.7 some descriptive statistics of elicited social preference parameters.

**Table 4.7.** *Elicited parameters of the other-regarding preference models*

| | | Theoretical Range | Empirical range | 25th percentile | Mean | 75th percentile | St. dev. |
|---|---|---|---|---|---|---|---|
| Inequality. Aversion | | | | | | | |
| | $\alpha_i$ | $[0, \infty)$ | [0,3] | 0.52 | 1.21 | 2.13 | 0.95 |
| | $\beta_i$ | $[0,1)$ | [0,1] | 0.05 | 0.38 | 0.55 | 0.35 |
| Spite | | | | | | | |
| | $\beta_i$ | $(-\infty, 0]$ | $[-.61,0]$ | 0.00 | -0.02 | 0.00 | 0.09 |
| Reciprocity | | | | | | | |
| | $\gamma_{i,-i}$ | $[0, \infty)$ | [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 cooperation 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 cooperation 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 cooperation 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 $\beta_i$. I imputed a value of 0 for the spite parameter to all subjects who revealed a positive $\beta_i$, 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.

### 4.5.2. *Do social preferences and moral rules influence cooperation attitudes?*

#### 4.5.2.1. *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 cooperation attitudes as the dependent variable and the predicted cooperation attitudes of most of the theories presented in the two previous sections as dependent variables[40]. Recall that cooperation 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 cooperation attitudes of a given subject in a game

[40] 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 since their predictions are perfectly correlated with the constant of regression. The analysis of 4.5.2.2. includes all the models in the comparison.

consist of four contributions, giving me a dependent variable with four observations per each subject for a given game. The predicted cooperation 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 cooperation attitudes of blame avoidance and praise seeking are calculated using the elicited moral judgments in the M-experiments (see subsections 4.3.2 and 4.3.3), the predicted cooperation 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 subsection 4.4 and Appendix C.2) 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 cooperation 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)[41]. Columns '*Estimates*' in Table 4.8 report the regression estimates.

Four patterns in the data reveal the role of each of the analysed theories in predicting cooperation 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 cooperation 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 cooperation attitudes. Third, reciprocity is statistically significant only in common interest games, suggesting that it is a specific motivation of cooperation 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,

[41] My rationale is as follows. First, note that guilt aversion's prediction, in social dilemmas, of cooperation 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 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 $c_{-i}$.

inequality aversion and reciprocity have a significantly greater coefficient in CIG, suggesting they play a greater role in explaining cooperation attitudes in CIG.

Additionally, I report the estimates of the decomposition of explained variance in columns ‘*Decomposition of* $R^2$’ of Table 4.8. 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 cooperation attitudes.

**Table 4.8.** *Regression estimates and decomposition of explained variance*

| **Dependent variable:** Cooperation 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.

It is remarkable to see that more than 50% of the explained variation in cooperation 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 cooperation 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 cooperation 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 cooperation attitudes in social dilemmas. Out of the remaining variables, only social efficiency has a non-negligible relative importance, although its role in explaining cooperation 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 cooperation attitudes of cooperation 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 cooperation attitudes in both cooperation problems. Second, reciprocity, social efficiency, and spite are game-specific explanations of cooperation 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 cooperation attitudes of social dilemmas, and social preferences play a greater role than moral rules in explaining cooperation attitudes of common interest games.

#### 4.5.2.2. *A revealed preference approach*

I complement the econometric analysis of the previous subsection with an additional one because of one main concern. 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.

In this complementary analysis, 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 cooperation 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.

#### 4.5.2.2.1. *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:

$$C(t, E^g, g) = \frac{n^g}{N} \tag{4.23}$$

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:

$$I(t, E^g, g) = \frac{o^g}{N} \tag{4.24}$$

The rationale for using these two ratios to analyse the data is as follows[42]. The degree of confirmation, or the hit rate, of a given theory captures the share of the total data for which a 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 cooperation 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.

[42] One can also trace the use of hit rates 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 Helmer and Oppenheim (1945, pp.50). Also, see Popper (2002, part 2, chapter 10, §79) for a critique of its use. Furthermore, one can interpret the degree of indubitable confirmation we propose to capture the share of what Bacon (2000, Book II, Aphorism XXXVI, pp.159-168) would call '*instantiae crucis*'.

#### 4.5.2.2.2. *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[43]. I present the estimates of $C(t,E^g,g)$ and $I(t,E^g,g)$ for all the theories I study in Table 4.9.

**Table 4.9.** *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.*

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 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

[43] 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.

the most probable explanations of cooperation 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 cooperation attitudes of CIG than of SDG. The evidence supports the inference of spite being decisively rejected as the explanation of cooperation attitudes in SDG and CIG.

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 cooperation in the SDG and unconditional cooperation 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.

## 4.6. Concluding remarks and implications of the results

In this chapter I have analysed the likelihood of a set of social preference and moral rule theories in explaining cooperation attitudes of two cooperation problems: social dilemmas and common interest games. To achieve this, I have measured (i) cooperation 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 cooperation attitudes in cooperation 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 cooperation attitudes) were to be true. In addition, I was able to decompose the results into each theory's share of explained variation of cooperation 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 cooperation 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 cooperation attitudes. In conclusion, cooperation attitudes of cooperation 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 cooperation 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 cooperation in the SDG and CIG and Dufwenberg and Kirchsteiger's (2004) theory predicts either non-perfect conditional cooperation in both games or free riding in the SDG and non-perfect conditional cooperation in the CIG or free riding in the SDG and unconditional cooperation in the CIG. 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 cooperation in both games, Dufwenberg and Kirchsteiger (2004) to capture non-perfect conditional cooperation in both games, maximin to capture perfect conditional cooperation in the SDG and unconditional cooperation in the CIG, social efficiency to capture unconditional cooperation in both the SDG and the CIG, and spite to capture free riding in the SDG and conditional cooperation 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 chapter 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 cooperation 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 cooperation 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.

# Chapter 5. Conclusions

The main objective of the thesis was to present a new framework of decision-making based on moral judgments. In the introduction I have contextualised where this project sits in between the dispute as to whether judgments or normative aspects of ethics can have a space in positive economics. I have enquired on whether moral judgments do in fact matter for people when making decisions. The findings of Chapter 2 showed that public goods are, indeed, seen as of moral significance, and Chapters 3 and 4 have demonstrated, not only that they are able to predict behaviour in public goods (Chapter 3), but also that they add explanatory power on top of classical theories of other-regarding preferences developed in the last four decades (Chapter 4). Hence, this thesis provides support for Hume's conjecture that morals are part of the *practical* nature of human beings. This raises important questions. For instance, would Friedman concede that normative concepts may have a role within the scope of positive economics, when moral judgments predict a certain behaviour that models not incorporating them cannot predict? And, if so, where do we draw the line between positive and normative economics – or should we directly erase that line? These are questions that go beyond what this thesis is about, but what is clear is that the results of the core chapters of the thesis support the view of Alfred Marshall when he wrote that '*ethical forces*' should be among the ones that an economist should consider in the explanation of socioeconomic phenomena.

Not only have I found support for the role of ethics in economic behaviour, but I have presented two specific models of moral decision-making, embedded within a general framework that I expect to develop further in the years to come. Additionally, I have provided a specific way to falsify the theory by (i) generating a new elicitation method on how to measure moral judgments of all strategy combinations of a given game, based on the methods of Cubitt et al (2011); and (ii) providing a quantitative structure to the theory that is able to make precise, testable predictions about cooperative behaviour. The structure is generic enough to be applied to a wider range of phenomena, which would be the next natural steps of this work.

Chapter 2 is the first substantial chapter of the thesis, providing some data on how subjects perceive a person's play in two public goods games (a give-some and a take-some version) from an impartial spectator viewpoint, that is, from a position where he/she has no stakes in the decision situation. I presented subjects with some scenarios, where each scenario provided a description of the decision situation, the contribution of the judged person ($C_A$), the average contributions of the non-judged group members ($C_{BC}$), the framing of the decision situation ($f$) and whether the contributions of the non-judged members where equal or different ($d$). My findings, in short, are as follows. On average, (i) subjects perceive public goods problems as morally significant, as the average moral judgments of scenarios differ significantly from 0; (ii) subjects' praise of Person A is increasing in $C_A$; (iii) subjects' Moral Evaluation Functions, as defined in Chapters 2, 3, and 4, are decreasing in $C_{BC}$; (iv) the negative slope of the MEF's becomes flatter the higher $C_A$; (v) subjects' perceptions of effective free riders (resp. effective half contributors) are more condemnatory (resp. more praiseworthy) in give-some than in take-some public goods problems, and effective full contributors are regarded as equally praiseworthy; and (vi) the dispersion in the contribution of the non-judged group members ($d = Unequal$ with respect to $d = Equal$) underemphasizes the condemnation (resp. praise) of effective free riders (resp. effective half contributors). Additionally, I find that the moral foundations of either Moral Foundations Theory or Morality As Cooperation Theory influence the moral judgments of subjects, though their role is significantly smaller than either of the manipulated variables commented earlier on.

Chapter 3 applied the design of Fischbacher and Gächter (2010), along with an elicitation of the moral judgments of all strategy combinations, to two cooperation problems: those of Provision and Maintenance of public goods (or give-some and take-some situations). I (i) investigated whether blame avoidance and praise seeking were motivations underlying contribution attitudes of Provision and Maintenance problems, controlling for the contribution of the other group member; and (ii) studied the explanatory power of blame avoidance and praise seeking along with the ABC approach, that maps contribution attitudes to unconditional contributions using subjects' beliefs about the contribution of the other group member, with respect to unconditional contributions to public goods. I found that blame avoidance, and not praise seeking, drive cooperation attitudes on both Provision and Maintenance problems; and that both blame avoidance, praise-seeking and the ABC method are significant drivers of unconditional contributions to public goods. I divided, through a mediation analysis, the effects that blame avoidance and praise seeking (through predicting contribution attitudes that feed into ABC) have in unconditional contributions, and I found evidence of substantial indirect effects. Perhaps as importantly, I find that blame avoidance and praise seeking are more important motives in explaining unconditional contributions in Maintenance problems relative to their role in explaining Provision problems.

Finally, Chapter 4 made a more formal presentation of the blame avoidance and praise seeking theories through what I called the MRC framework and brought the theories in a more stringent test against six models of decision making: material selfishness, inequality aversion, reciprocity, social efficiency, maximin, and spite. I used contribution attitudes to a social dilemma and a common interest game as the environment where to perform the test of the theories, allowing for an *ex-ante* theoretical separation in the predictions of the joint play in both games. I found evidence that all theories, except spite, play a role in contribution attitudes, blame avoidance, inequality aversion and maximin being the most important factors driving contribution attitudes in both games; and social efficiency, reciprocity, and praise seeking theories being predictors of contribution attitudes mostly towards common interest games.

Overall, the results provide evidence for the following claims. First, some economic issues (cooperation problems) do not only have economic implications, or social implications such as inequality, social efficiency, or unkindness. Rather, how people choose to play in the game has moral implications on how people perceive those persons. Second, impartial moral judgments are important drivers of cooperative attitudes beyond the scope of what can be captured by canonical models of other-regarding preferences. Third, a tendency to avoid doing what one perceives as blameworthy, and a tendency to approach doing what one perceives as praiseworthy from an impartial perspective, are important considerations that some subjects consider in their cooperative choices.

# Chapter 6. Appendices

## Appendix A

### *A.1. Instructions of Study 1 and Study 2*

Below we provide the instructions of Study 1 and Study 2. Parts 1 to 3 presented below are the same in Study 1 and Study 2. Part 4 – the moral questionnaire – differs between studies. We present both moral questionnaires used (MFQ and MAC-Q) and make clear which one corresponds to which study. Parts 1 and 2 vary according to the frame. We write in brackets and italics the different parts of the text between the give and the take treatments. Part 3 is the sociodemographic questionnaire and was always presented between part 2 and part 4. We preserve the format of the text as presented to the subjects to the best of our abilities. Qsf (Qualtrics) files for both studies will be provided upon request.

Part 1 – Introduction to the give [*take*] decision situation and control questions

# Please read through the decision situation outlined below and answer a set of questions related to it.

Imagine a group that consists of three people: Person A, Person B and Person C.

The three group members share a common project. Initially, the project has 0 [*60*] tokens and each group member has 20 [*0*] tokens in a private account. Each group member must decide how many tokens to contribute to [withdraw from] the project. They only have three options: either contribute [*withdraw*] 0, 10 or 20 tokens to [*from*] the project. Tokens contributed [*withdrawn*] are transferred to the project [*group member's own private account*]. Tokens not contributed [*withdrawn*] are retained in the group member's own private account [*project*]. All three group members face the same options.

Each group member receives an income from their private account and from the project.

# Their income from the private account

**Each group member receives $1 for each token they retain in their own private account [*withdraw from the project*].** For example, if they decide to retain [*withdraw*] 20 tokens in their private account [*from the project*], their income from their private account will be $20. If they decide to retain [*withdraw*] 10 tokens in their private account [*from the project*], their income from their private account will be $10. **No one except person A receives anything from tokens person A retain in his or her private account [*withdraws from the project*]. The same holds for the other two group members.**

# Their income from the project

**Each group member benefits equally from what any group member contributes to [*retains in*] the project.** All tokens contributed to [*retained in*] the project are converted into dollars, **increased by 50 percent (i.e., multiplied by a factor of 1.5) and split equally** among the three group members. That is, for every token contributed to [*retained in*] the project by any group member, each of the three group members receives: **$1 × 1.5 / 3 = $0.5**.

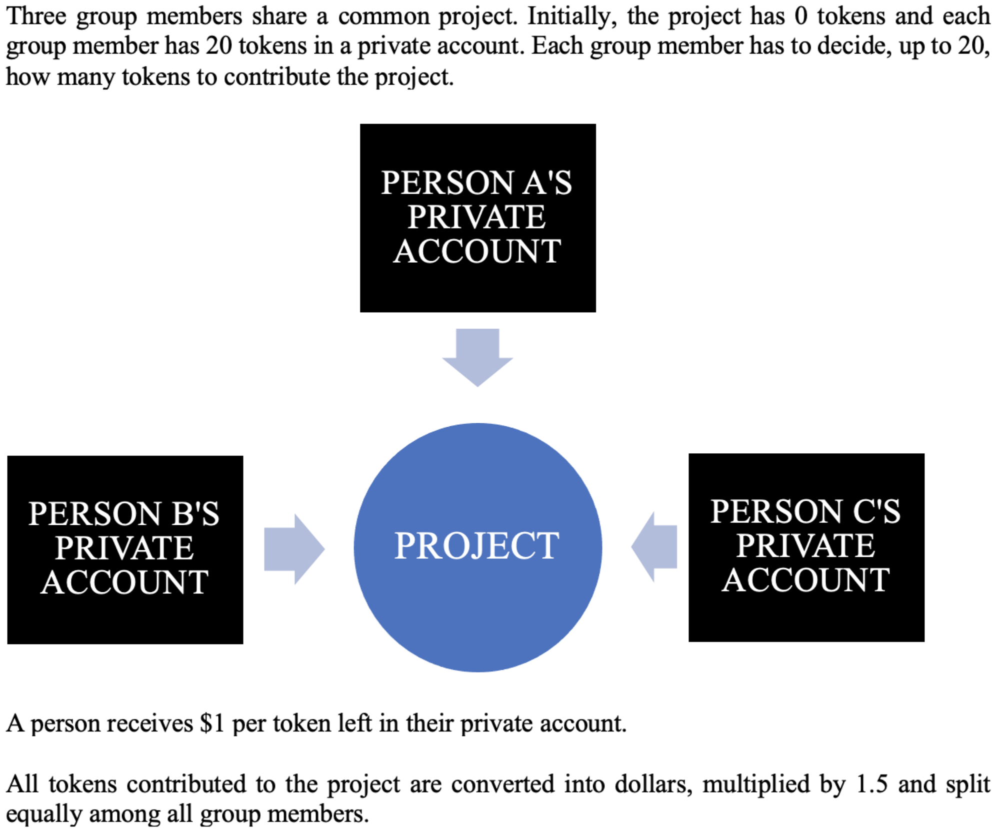
If, for example, the sum of all tokens contributed to [*retained in*] the project by the three group members is 60 tokens, then each group member receives $**60 × 1.5 / 3 = 60 × 0.5 = $30** from the project.

If, for example, the sum of all tokens contributed to [*retained in*] the project by the three group members is 10 tokens, then each group member receives $**10 × 1.5 / 3 = 10 × 0.5 = $5** from the project.

## Their total income

Each member's total income is the sum of the income they receive from their private account and the income they receive from the project. The figure below shows a summary of the interaction:

*[Here, the first figure corresponds to what is shown in the give treatments and the second figure corresponds to what is displayed in the take treatments]*



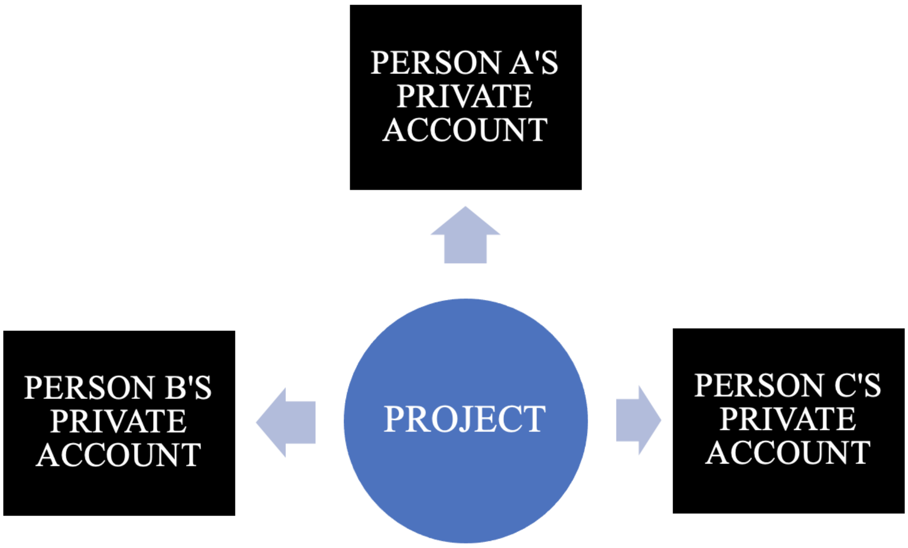
Three group members share a common project. Initially, the project has 0 tokens and each group member has 20 tokens in a private account. Each group member has to decide, up to 20, how many tokens to contribute the project.

PERSON A'S PRIVATE ACCOUNT
PERSON B'S PRIVATE ACCOUNT
PROJECT
PERSON C'S PRIVATE ACCOUNT

A person receives $1 per token left in their private account.

All tokens contributed to the project are converted into dollars, multiplied by 1.5 and split equally among all group members.

Three group members share a common project. Initially, the project has 60 tokens and each group member has 0 tokens in a private account. Each group member has to decide, up to 20, how many tokens to withdraw from the project.


PERSON A'S PRIVATE ACCOUNT
PERSON B'S PRIVATE ACCOUNT
PROJECT
PERSON C'S PRIVATE ACCOUNT

A person receives $1 per token withdrawn from the project.

All tokens retained in the project are converted into dollars, multiplied by 1.5 and split equally among all group members.

**Please answer the following questions to check your understanding of the situation.**

**If you answer any of the questions incorrectly you will be asked to reconsider the wrong answers.**

**Q1.** Assume that all three group members (including person A) withdraw 20 tokens each from the project.

What are the total earnings (in dollars) of person A, person B and person C (= earnings from the private account + earnings from the project)?

Person A earnings ________
Person B earnings ________
Person C earnings ________
**Q2.** Assume that all three group members (including person A) withdraw 0 tokens each from the project.

What are the total earnings (in dollars) of person A, person B and person C (= earnings from the private account + earnings from the project)?

Person A earnings ________
Person B earnings ________
Person C earnings ________

**Q3.** Assume person A withdraws 20 tokens from the project and the other group members withdraw 0 tokens each from the project.

What are the total earnings (in dollars) of person A, person B and person C (= earnings from the private account + earnings from the project)?

Person A earnings ________
Person B earnings ________
Person C earnings ________

**Q4.** Assume person A withdraws 0 tokens from the project, person B withdraws 0 tokens from the project and person C withdraws 20 tokens from the project.

What are the total earnings (in dollars) of person A, person B and person C (= earnings from the private account + earnings from the project)?

Person A earnings ________
Person B earnings ________
Person C earnings ________

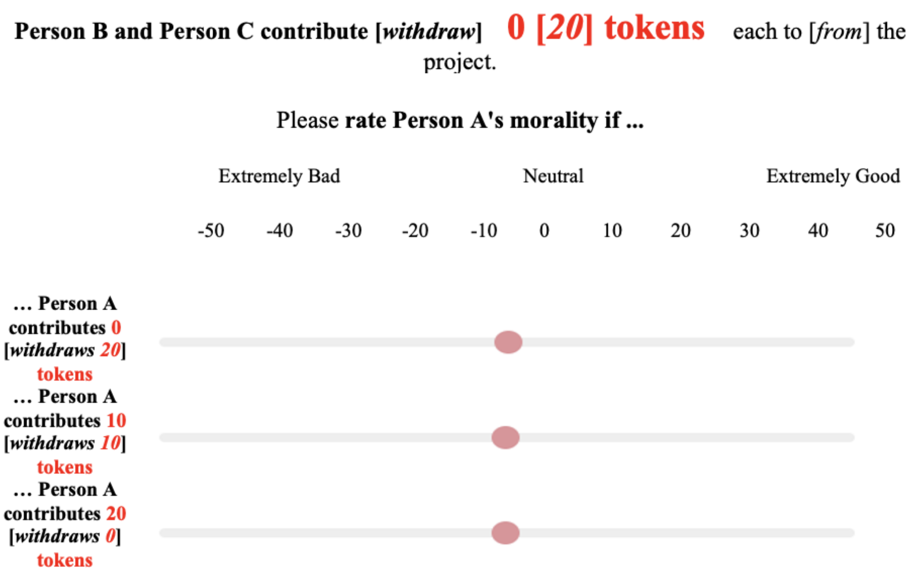
*[Subjects had to answer all the questions correctly to proceed to the next parts of the study. Otherwise, they were not allowed to complete the study]*

Part 2 – Moral Evaluation of the Scenarios

# Thank you for finishing the previous part. Now, please rate the morality of Person A in several different scenarios.

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.

*[We displayed the scenarios in 4 different screens, three scenarios per each screens. The three scenarios per screen fixed what Person B and Person C did in a line summarising their actions. Each scenario per screen varied what Person A did. The scenarios displayed below have been worked so that they resemble how the screens looked to the participant]*



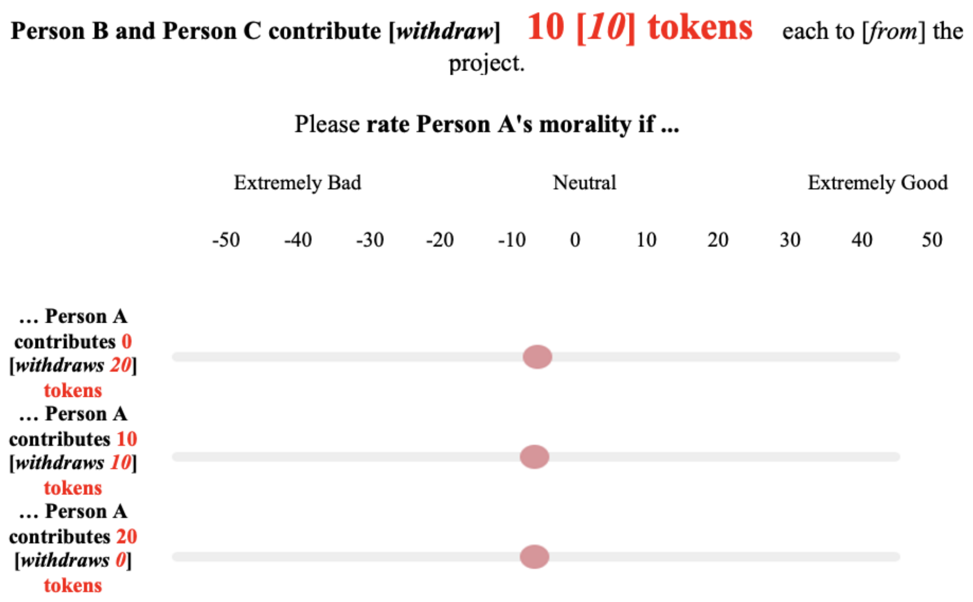
**Person B and Person C contribute [*withdraw*]** **0 [*20*] tokens** each to [*from*] the project.

Please **rate Person A's morality if ...**

**... Person A contributes 0 [*withdraws 20*] tokens**

**... Person A contributes 10 [*withdraws 10*] tokens**

**... Person A contributes 20 [*withdraws 0*] tokens**



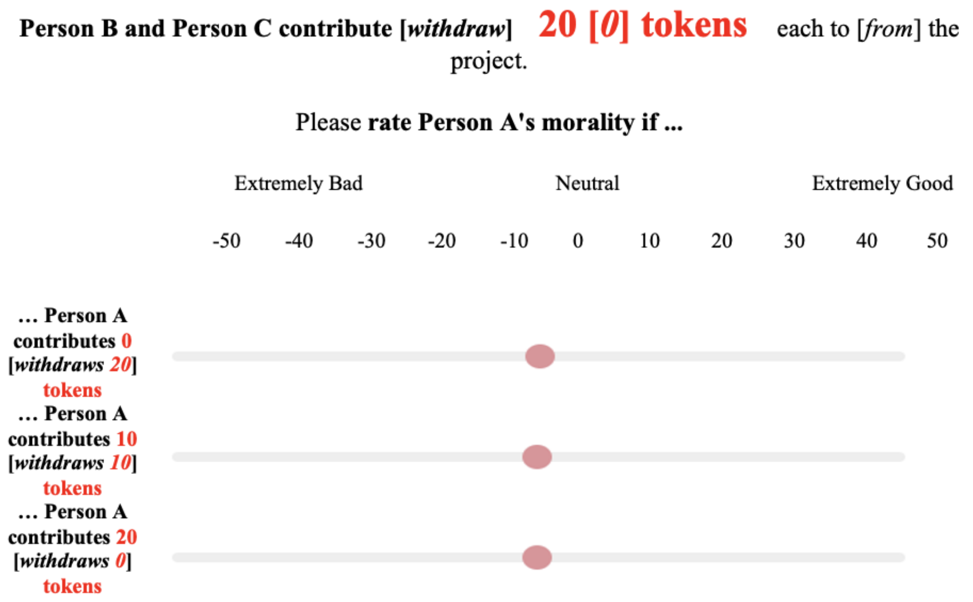
**Person B and Person C contribute [*withdraw*] 10 [*10*] tokens** each to [*from*] the project.

Please **rate Person A's morality if ...**

**... Person A contributes 0 [*withdraws 20*] tokens**

**... Person A contributes 10 [*withdraws 10*] tokens**

**... Person A contributes 20 [*withdraws 0*] tokens**



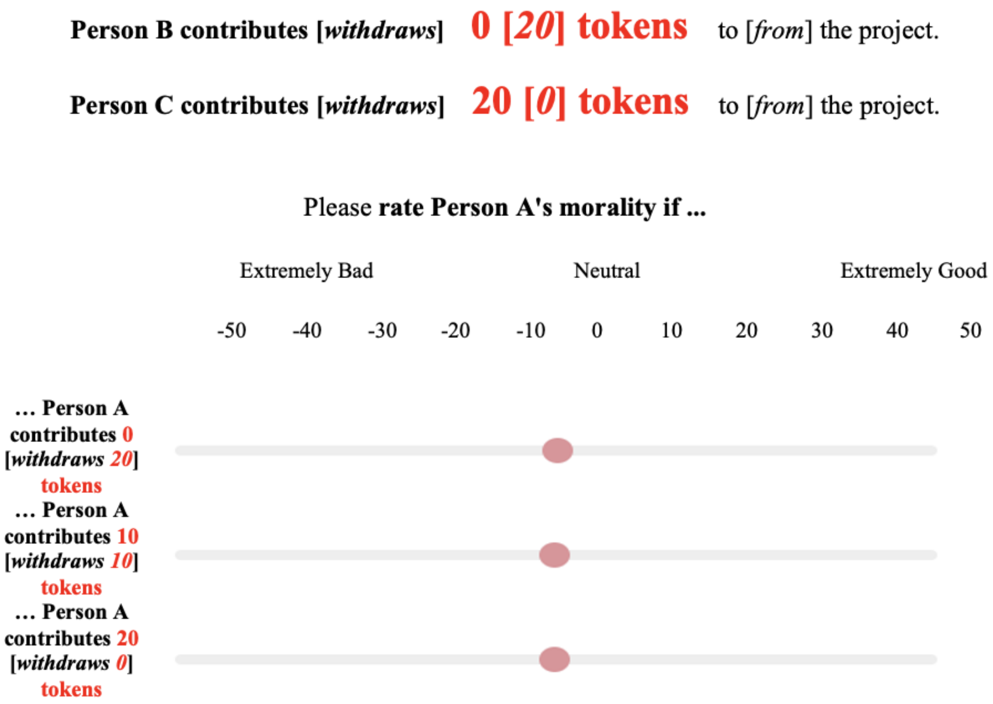
**Person B and Person C contribute [*withdraw*] 20 [*0*] tokens** each to [*from*] the project.

Please **rate Person A's morality if ...**

**... Person A contributes 0 [*withdraws 20*] tokens**

**... Person A contributes 10 [*withdraws 10*] tokens**

**... Person A contributes 20 [*withdraws 0*] tokens**



**Person B contributes [*withdraws*]** **0 [*20*] tokens** to [*from*] the project.

**Person C contributes [*withdraws*]** **20 [*0*] tokens** to [*from*] the project.

Please **rate Person A's morality if ...**

**… Person A contributes 0 [*withdraws 20*] tokens**

**… Person A contributes 10 [*withdraws 10*] tokens**

**… Person A contributes 20 [*withdraws 0*] tokens**

Part 3 – Sociodemographic questionnaire

**Thank you for finishing the previous part. Now, please read through the questions below and answer them as accurately as possible.**

*[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.** How many hours in total do you work per week?
*[Options to the respondent: 0 – 20, 20 – 40, 40 – 60, 60 – 80, More than 80]*

**Q2.** Your Gender:
*[Options to the respondent: Male, Female, Prefer not to say]*

**Q3.** Your Age:
*[Options to the respondent: from 15 to 100 in steps of 1]*

**Q4.** What is your nationality?
*[Options to the respondent: The default list of countries]*

**Q5.** Would you describe yourself as a liberal, conservative or something else?

*[Options to the respondent: Moderate/Middle of the Road, Liberal, Very Liberal, Conservative, Very Conservative, Libertarian, Other, Prefer not to say]*

**Q6.** How religious are you?

*[Options to the respondent: Not at all, Somewhat religious, Very religious, Prefer not to say]*

**Q7.** 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]*

**Q8.** What is your highest qualification attained?

*[Options to the respondent: Less than high school, High school, Vocational Training, Attended University but didn't finish, Undergraduate Degree, Postgraduate Degree, Prefer not to say]*

**Q9.** Please choose the category that describes the total amount of income you earned this year.

*[Options to the respondent: $5,000 or less, $5,001 – $25,000, $25,001 – $50,000, $50,001 – $75,000, $75,001 – $100,000, More than $100,000, Prefer not to say]*

**Q10.** 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]*

Part 4 – Moral questionnaire

## Study 1 – Moral Foundations Questionnaire (MFQ)

**Thank you for finishing the previous part. Now, please answer the following questionnaire.**

**Part 1.** When you decide whether something is right or wrong, to what extent are the following considerations relevant to your thinking? Please answer on a scale from

*Not At All Relevant* (this consideration has nothing to do with my judgments of right and wrong)

to *Extremely Relevant* (this is one of the most important factors when I judge right and wrong)

*[Options for the respondent: Not At All Relevant, Not Very Relevant, Slightly Relevant, Somewhat Relevant, Very Relevant, Extremely Relevant]*

*[Each sentence was displayed with Font Times New Roman, size 18, bold and centred. The options for the respondent appeared below each of the statements]*

Whether or not someone conformed to the traditions of society.
Whether or not an action caused chaos or disorder.
Whether or not some people were treated differently than others.
Whether or not someone was denied his or her rights.
Whether or not someone acted in a way that God would approve of.
Whether or not someone showed a lack of respect for authority.
Whether or not someone was good at math.
Whether or not someone's action showed love for his or her country.
Whether or not someone cared for someone weak or vulnerable.
Whether or not someone suffered emotionally.

Whether or not someone showed a lack of loyalty.

Whether or not someone did something disgusting.

Whether or not someone did something to betray his or her group.

Whether or not someone acted unfairly.

Whether or not someone violated standards of purity and decency.

Whether or not someone was cruel.

**Part 2.** Please read the following sentences and indicate your level of agreement or disagreement

*[Options for the respondent: Strongly Disagree, Moderately Disagree, Slightly Disagree, Slightly Agree, Moderately Agree, Strongly Agree]*

*[Each sentence was displayed with Font Times New Roman, size 18, bold and centred. The options for the respondent appeared below each of the statements]*

It is more important to be a team player than to express oneself.

Men and Women have different roles to play in society.

Chastity is an important and valuable virtue.

I am proud of my country's history.

Justice is the most important requirement for a society.

I would call some acts wrong on the grounds that they are unnatural.

Compassion for those who are suffering is the most crucial virtue.

It is better to do good than to do bad.

One of the worst things a person could do is hurt a defenseless animal.

People should be loyal to their family members, even when they have done something wrong.

When the government makes laws, the number one principle should be ensuring that everyone is treated fairly.

I think it's morally wrong that rich children inherit a lot of money while poor children inherit nothing.

Respect for authority is something all children need to learn.

People should not do things that are disgusting, even if no one is harmed.

If I were a soldier and disagreed with my commanding officer's orders, I would obey anyway because that is my duty.

It can never be right to kill a human being.

# Study 2 – Morality As Cooperation Questionnaire (MAC-Q)

**Thank you for finishing the previous part. Now, please answer the following questionnaire.**

*[MAC-Q is displayed differently (and on a different scale) than the MFQ. We preserve how MAC-Q was displayed in Curry et al (2019). To do that, we displayed each of the statements in a slider frame (as the moral evaluations of scenarios)]*

**When you decide whether something is right or wrong, to what extent are the following considerations relevant to your thinking?**

Click on the lines below, and/or move the sliders

*[Options for the respondent: Continuous scale via slider. Guide above the sliders showed the following guide: Not At All Relevant, Not Very Relevant, Slightly Relevant, Somewhat Relevant, Very Relevant, Extremely Relevant]*

Whether or not someone acted to protect their family

Whether or not someone helped a member of their family

Whether or not someone’s action showed love for their family

Whether or  not someone acted in a way that helped their community

Whether or not someone helped a member of their community

Whether or not someone worked to unite a community

Whether or not someone did what they had agreed to do

Whether or not someone kept their promise

Whether or not someone proved that they could be trusted

Whether or not someone acted heroically

Whether or not someone showed courage in the face of adversity

Whether or not someone was brave

Whether or not someone deferred to those in authority

Whether or not someone disobeyed orders

Whether or not someone showed respect for authority

Whether or not someone kept the best part for themselves

Whether or not someone showed favouritism

Whether or not someone took more than others

Whether or not someone vandalised another person's property

Whether or not someone kept something that didn't belong to them

Whether or not someone's property was damaged

## To what extent do you agree with the following statements?

Click on the lines below, and/or move the sliders

*[Continuous scale via slider. Guide above the sliders showed the following guide: Strongly Disagree, Disagree, Neither Agree nor Disagree, Agree, Strongly Agree]*

People should be willing to do anything to help a member of their family

You should always be loyal to your family

You should always put the interests of your family first

People have an obligation to help members of their community

It's important for individuals to play an active role in their communities

You should try to be a useful member of society

You have an obligation to help those who have helped you

You should always make amends for the things you have done wrong

You should always return a favour if you can

Courage in the face of adversity is the most admirable trait

Society should do more to honour its heroes

To be willing to lay down your life for your country is the height of bravery

People should always defer to their superiors

Society would be better if people were more obedient to authority

You should respect people who are older than you

Everyone should be treated the same

Everyone’s rights are equally important

The current levels of inequality in society are unfair

It’s acceptable to steal food if you are starving

It’s ok to keep valuable items that you find, rather than try to locate the rightful owner

Sometimes you are entitled to take things you need from other people

### *A.2. Replication of the main empirical regularities of the Foundations in MFT and MAC*

The tables provided in this appendix serve the purpose of supporting the claim that the data we collected from the Moral Foundations Questionnaire (Study 1. Henceforth, MFQ) and the Morality As Cooperation Questionnaire (Study 2. Henceforth, MAC-Q) are in line with the data that has been reported in the previous literature (most notably, Graham et al, 2011 and Curry et al, 2019, which are the papers who developed and presented MFWQ and MAC-Q respectively). We present several tables for each foundation, divided into two sections: one for MFQ (A.2.1) and one for MAC-Q (A.2.2). In each section, our tables present (i) the Cronbach alphas of each foundation, aggregated over all measures and divided by Relevance and Judgment items; (ii) correlations (both before and after partialing political ideology) between the Relevance and Judgments subscales for each Foundation; (iii) the Factor Loadings from an Exploratory Factor Analysis for both the Judgment and Relevance items of each foundation; and (iv) the goodness-of-fit indices of several modelling structures of MFT and MAC-Q used in Confirmatory Factor Analyses.

In summary, we replicate very well, with minor exceptions, the data from Graham et al (2011) for MFT and the data from, Curry et al (2019) for MAC, albeit the former with greater precision. More specifically, we can summarise the results into four statements. First, the Cronbach alphas of our data resemble those obtained in Graham et al (2011) and Curry et al (2019), both in size and in the fact that Cronbach alphas are systematically higher for the Relevance than for the Judgments items. Second, the correlation between the Relevance items of a given foundation tend to correlate more strongly with the Judgment items of the same foundation, both in Graham et al (2011), Curry et al (2019) and in our data for Study 1 and Study 2. Third, the Factor loadings of MFQ can be divided into a factor for Harm and Fairness and a factor for Loyalty, Authority, and Purity both in Graham et al (2011) and in our data of Study 1. This replication is still successful for our data of Study 2, but only for the Relevance items, in which we had 7 factors, one per each Foundation, as Curry et al (2019). Our factor loadings for the Judgments items, however, are the only part of this analysis where we fail to replicate Curry et al's (2019) Studies. Fourth, the goodness-of-fit indices from the Confirmatory Factor Analyses favour in our data the same models as the ones favoured in the data of Graham et al (2011) and Curry et al (2019). Overall, we take this as strong evidence that our data is representative of what has already been reported in the literature, hence giving more credibility to the findings we report regarding the influence of Moral Foundations in capturing individual differences in the moral judgments of our scenarios.

### *A.2.1. Moral Foundations Theory*

**Table 6.1.** *Cronbach Alphas, means and standard deviations for each subscale of MFQ*

**Graham et al (2011)**

| Foundation | Subscale | $\alpha$ | Total | Liberals | Moderates | Conservatives |
|---|---|---|---|---|---|---|
| Harm | Relevance | 0.70 | 3.77 (0.86) | 3.93 (0.76) | 3.68 (0.84) | 3.48 (0.89) |
| | Judgments | 0.51 | 3.08 (1.11) | 3.32 (1.01) | 2.95 (1.09) | 2.48 (1.11) |
| | Total | 0.69 | 3.42 (0.84) | 3.62 (0.74) | 3.31 (0.81) | 2.98 (0.84) |
| Fairness | Relevance | 0.65 | 3.89 (0.78) | 4.04 (0.67) | 3.77 (0.77) | 3.44 (0.87) |
| | Judgments | 0.40 | 3.21 (0.93) | 3.43 (0.86) | 3.00 (0.86) | 2.59 (0.87) |
| | Total | 0.65 | 3.55 (0.73) | 3.74 (0.63) | 3.39 (0.68) | 3.02 (0.73) |
| Loyalty | Relevance | 0.71 | 2.24 (1.03) | 2.06 (0.94) | 2.56 (1.00) | 3.03 (1.02) |
| | Judgments | 0.46 | 2.28 (0.98) | 2.09 (0.91) | 2.59 (0.90) | 3.13 (0.85) |
| | Total | 0.71 | 2.26 (0.87) | 2.07 (0.77) | 2.58 (0.79) | 3.08 (0.79) |
| Authority | Relevance | 0.67 | 2.03 (0.95) | 1.88 (0.86) | 2.37 (0.90) | 2.81 (0.91) |
| | Judgments | 0.60 | 2.52 (1.12) | 2.23 (1.01) | 2.97 (0.94) | 3.74 (0.82) |
| | Total | 0.74 | 2.27 (0.90) | 2.06 (0.79) | 2.67 (0.77) | 3.28 (0.71) |
| Purity | Relevance | 0.68 | 1.68 (1.11) | 1.44 (0.94) | 2.09 (1.09) | 2.88 (1.11) |
| | Judgments | 0.75 | 1.41 (1.20) | 1.09 (0.96) | 1.88 (1.16) | 2.90 (1.20) |
| | Total | 0.84 | 1.54 (1.08) | 1.27 (0.86) | 1.99 (1.03) | 2.89 (1.07) |

**Study 1**

| Foundation | Subscale | $\alpha$ | Total | Liberals | Moderates | Conservatives |
|---|---|---|---|---|---|---|
| Harm | Relevance | 0.75 | 3.88 (0.83) | 4.05 (0.78) | 3.81 (0.86) | 3.57 (0.85) |
| | Judgments | 0.46 | 3.55 (0.97) | 3.77 (0.91) | 3.31 (0.92) | 3.33 (0.99) |
| | Total | 0.71 | 3.71 (0.78) | 3.91 (0.72) | 3.56 (0.76) | 3.45 (0.83) |
| Fairness | Relevance | 0.74 | 4.03 (0.78) | 4.2 (0.73) | 4.04 (0.69) | 3.61 (0.86) |
| | Judgments | 0.16 | 3.38 (0.83) | 3.64 (0.78) | 3.27 (0.68) | 2.94 (0.83) |
| | Total | 0.57 | 3.7 (0.66) | 3.92 (0.6) | 3.66 (0.58) | 3.28 (0.7) |
| Loyalty | Relevance | 0.79 | 2.13 (1.07) | 1.99 (1.06) | 2.01 (1.03) | 2.66 (1.02) |
| | Judgments | 0.49 | 2.24 (0.98) | 1.99 (0.87) | 2.22 (0.98) | 3 (0.84) |
| | Total | 0.74 | 2.19 (0.88) | 1.99 (0.82) | 2.12 (0.87) | 2.83 (0.77) |
| Authority | Relevance | 0.73 | 2.4 (1.03) | 2.22 (1.02) | 2.39 (0.94) | 2.98 (0.94) |
| | Judgments | 0.66 | 2.66 (1.16) | 2.33 (1.1) | 2.7 (1.02) | 3.6 (0.87) |
| | Total | 0.79 | 2.53 (0.98) | 2.28 (0.95) | 2.55 (0.86) | 3.29 (0.73) |
| Purity | Relevance | 0.74 | 1.89 (1.31) | 1.56 (1.23) | 1.88 (1.28) | 2.85 (1.01) |
| | Judgments | 0.79 | 2.06 (1.39) | 1.61 (1.21) | 2.06 (1.43) | 3.26 (1.03) |
| | Total | 0.88 | 1.98 (1.28) | 1.59 (1.15) | 1.97 (1.29) | 3.06 (0.93) |

*Notes*: Range for all subscales is 0–5. Standard deviations in parentheses.

**Table 6.2.** *Correlations between the Relevance and Judgments subscales of MFQ*

**Graham et al (2011)**

| | MFQ Relevance subscales | | | | |
|---|---|---|---|---|---|
| | Harm | Fairness | Loyalty | Authority | Purity |
| MFQ Judgments subscales | | | | | |
| Harm | **0.47** | 0.36 | 0.03 | 0.04 | 0.10 |
| Fairness | 0.32 | **0.46** | - 0.09 | - 0.11 | - 0.12 |
| Loyalty | - 0.05 | - 0.13 | **0.48** | 0.43 | 0.40 |
| Authority | - 0.12 | - 0.21 | 0.42 | **0.49** | 0.47 |
| Purity | 0.05 | - 0.09 | 0.44 | 0.53 | **0.74** |
| After partialing political ideology | | | | | |
| Harm | **0.38** | 0.25 | 0.07 | 0.09 | 0.19 |
| Fairness | 0.23 | **0.35** | 0.06 | 0.05 | 0.07 |
| Loyalty | 0.04 | - 0.04 | **0.38** | 0.32 | 0.25 |
| Authority | - 0.02 | - 0.08 | 0.31 | **0.39** | 0.31 |
| Purity | 0.11 | - 0.01 | 0.29 | 0.37 | **0.64** |

**Study 1**

| | MFQ Relevance subscales | | | | |
|---|---|---|---|---|---|
| | Harm | Fairness | Loyalty | Authority | Purity |
| MFQ Judgments subscales | | | | | |
| Harm | **0.50** | 0.30 | 0.08 | 0.08 | 0.11 |
| Fairness | 0.30 | **0.36** | -0.07 | -0.01 | -0.07 |
| Loyalty | -0.13 | -0.18 | **0.48** | 0.46 | 0.41 |
| Authority | -0.14 | -0.19 | 0.46 | **0.60** | 0.54 |
| Purity | -0.12 | -0.20 | 0.51 | 0.60 | **0.81** |
| After partialing political ideology | | | | | |
| Harm | **0.47** | 0.25 | 0.13 | 0.12 | 0.16 |
| Fairness | 0.24 | **0.30** | -0.02 | 0.03 | 0.00 |
| Loyalty | -0.07 | -0.13 | **0.47** | 0.45 | 0.39 |
| Authority | -0.07 | -0.13 | 0.45 | **0.59** | 0.52 |
| Purity | -0.05 | -0.14 | 0.51 | 0.60 | **0.80** |

*Notes*: Each panel shows correlations between scales (first five rows with data) and partial correlations after controlling for political ideology (last five rows with data). Highest correlation is shown in bold.

**Table 6.3.** *Factor loadings from Exploratory Factor Analysis for the MFQ items*

| | **Graham et al (2011)** | | **Study 1** | |
|---|---|---|---|---|
| Foundation | Factor 1 | Factor 2 | Factor 1 | Factor 2 |
| Relevance Items | | | | |
| Harm | | | | |
| Emotionally | 0.01 | **0.59** | 0.01 | **0.68** |
| Weak | 0.09 | **0.65** | -0.02 | **0.69** |
| Cruel | 0.07 | **0.59** | -0.03 | **0.59** |
| Fairness | | | | |
| Unfairly | 0.01 | **0.56** | -0.06 | **0.58** |
| Treated | - 0.11 | **0.59** | -0.06 | **0.72** |
| Rights | - 0.18 | **0.47** | -0.13 | **0.51** |
| Loyalty | | | | |
| Loyalty | **0.52** | 0.19 | **0.61** | 0.12 |
| Betray | **0.48** | 0.17 | **0.59** | 0.14 |
| Love countryok | **0.67** | 0.02 | **0.66** | 0.05 |
| Authority | | | | |
| Traditions | **0.61** | 0.00 | **0.70** | -0.04 |
| Respect | **0.69** | 0.05 | **0.81** | 0.08 |
| Chaos | **0.40** | 0.20 | **0.43** | 0.27 |
| Purity | | | | |
| Disgusting | **0.57** | 0.21 | **0.69** | 0.09 |
| Decency | **0.70** | 0.10 | **0.76** | 0.07 |
| God | **0.64** | - 0.02 | **0.61** | -0.07 |
| Judgment Items | | | | |
| Harm | | | | |
| Animal | - 0.01 | **0.39** | 0.05 | **0.31** |
| Kill | - 0.07 | **0.35** | 0.11 | **0.29** |
| Compassion | - 0.01 | **0.63** | -0.01 | **0.68** |
| Fairness | | | | |
| Justice | 0.09 | **0.27** | **0.20** | 0.11 |
| Fairly | - 0.14 | **0.48** | -0.10 | **0.56** |
| Rich | - 0.22 | **0.38** | -0.24 | **0.27** |
| Loyalty | | | | |
| Team | **0.45** | - 0.03 | **0.35** | -0.09 |
| History | **0.49** | - 0.18 | **0.47** | -0.21 |
| Family | **0.34** | 0.04 | **0.43** | -0.01 |
| Authority | | | | |
| Sex roles | **0.47** | - 0.24 | **0.50** | -0.32 |
| Soldier | **0.48** | - 0.20 | **0.49** | -0.07 |
| Kid respect | **0.64** | - 0.05 | **0.67** | 0.00 |
| Purity | | | | |
| Harmless disgusting | **0.66** | 0.03 | **0.74** | -0.11 |
| Unnatural | **0.66** | - 0.07 | **0.77** | -0.13 |
| Chastity | **0.67** | - 0.08 | **0.62** | -0.12 |

*Notes:* Strongest factor loading for each item indicated in bold.

**Table 6.4.** *Goodness-of-Fit Indices from Structural Models Representing Confirmatory Factor Analyses of MFQ*

| | | | **Relevance items** | | | | | |
|---|---|---|---|---|---|---|---|---|
| Model | $\chi^2$ | $df$ | RMSEA | AIC | BIC | CFI | TLI | SRMR |
| | | | *Graham et al (2011)* | | | | | |
| 1. Single factor | 57,093.30 | 90 | 0.10 | | | | | |
| 2. Two correlated factors | 20,180.00 | 89 | 0.06 | | | | | |
| **3. Five correlated factors** | 11,347.50 | 80 | 0.05 | | | | | |
| | | | *Curry et al (2019)* | | | | | |
| **3. Five correlated factors** | 482.13 | 80 | 0.07 | | | 0.90 | | 0.06 |
| | | | *Study 1* | | | | | |
| 1. Single factor | 12,826.53 | 90.00 | 0.18 | 190,988.16 | 191,274.35 | 0.54 | 0.46 | 0.16 |
| 2. Two correlated factors | 6,294.15 | 89.00 | 0.13 | 184,457.78 | 184,750.33 | 0.77 | 0.73 | 0.08 |
| **3. Five correlated factors** | 4,288.56 | 80.00 | 0.11 | 182,470.19 | 182,819.98 | 0.85 | 0.80 | 0.07 |
| | | | **Judgment items** | | | | | |
| Model | $\chi^2$ | $df$ | RMSEA | AIC | BIC | CFI | TLI | SRMR |
| | | | *Graham et al (2011)* | | | | | |
| 1. Single factor | 32,485.50 | 90 | 0.08 | | | | | |
| 2. Two correlated factors | 17,270.00 | 89 | 0.06 | | | | | |
| **3. Five correlated factors** | 11,084.60 | 80 | 0.05 | | | | | |
| | | | *Curry et al (2019)* | | | | | |
| **3. Five correlated factors** | 459.07 | 80 | 0.07 | | | 0.84 | | 0.06 |
| | | | *Study 1* | | | | | |
| 1. Single factor | | 90 | | 216,951.61 | 217,231.44 | | | 0.10 |
| 2. Two correlated factors | 5,251.32 | 89 | 0.12 | 215,465.76 | 215,758.31 | 0.70 | 0.65 | 0.10 |
| **3. Five correlated factors** | 4,301.13 | 80 | 0.11 | 214,533.57 | 214,883.36 | 0.76 | 0.68 | 0.09 |
| | | | **Full MFQ (all items)** | | | | | |
| Model | $\chi^2$ | $df$ | RMSEA | AIC | BIC | CFI | TLI | SRMR |
| | | | *Graham et al (2011)* | | | | | |
| 1. Single factor | 138,995.40 | 405 | 0.07 | | | | | |
| 2. Two correlated factors | 74,542.90 | 404 | 0.05 | | | | | |
| **3. Five correlated factors** | 53,894.10 | 395 | 0.05 | | | | | |
| | | | *Curry et al (2019)* | | | | | |
| 3. Five correlated factors | 2,393.62 | 395 | 0.07 | 91,863.99 | 92,041.27 | 0.76 | | 0.08 |
| **4. Different but related** | 1,553.76 | 360 | 0.06 | 90,909.77 | 91,149.10 | 0.86 | | 0.06 |
| | | | *Study 1* | | | | | |
| 1. Single factor | | 405 | | 404,331.25 | 404,897.27 | 1.00 | | 0.13 |
| 2. Two correlated factors | 20,778.44 | 404 | 0.11 | 394,089.05 | 394,667.80 | 0.66 | 0.63 | 0.09 |
| 3. Five correlated factors | 16,492.75 | 395 | 0.10 | 389,821.37 | 390,457.35 | 0.73 | 0.70 | 0.09 |
| **4. Different but related** | 13,502.35 | 360 | 0.09 | 386,900.97 | 387,759.55 | 0.78 | 0.73 | 0.08 |

*Notes:* Model in bold is the best fitting one according to the indices reported in here

### *A.2.2. Morality As Cooperation*

**Table 6.5.** *Cronbach Alphas, means and standard deviations for each subscale of MAC-Q*

**Curry et al (2019)**

| Foundation | Subscale | α | Total | Liberals | Moderates | Conservatives |
|---|---|---|---|---|---|---|
| Family | Relevance | 0.86 | 67.02 (18.56) | | | |
| | Judgments | 0.83 | 67.76 (17.50) | | | |
| | Total | | | | | |
| Group | Relevance | 0.86 | 59.77 (18.43) | | | |
| | Judgments | 0.75 | 64.64 (14.82) | | | |
| | Total | | | | | |
| Reciprocity | Relevance | 0.83 | 66.45 (18.01) | | | |
| | Judgments | 0.68 | 72.12 (12.71) | | | |
| | Total | | | | | |
| Heroism | Relevance | 0.84 | 61.84 (19.00) | | | |
| | Judgments | 0.71 | 66.17 (17.50) | | | |
| | Total | | | | | |
| Deference | Relevance | 0.80 | 53.89 (19.14) | | | |
| | Judgments | 0.69 | 54.71 (17.82) | | | |
| | Total | | | | | |
| Fairness | Relevance | 0.76 | 56.47 (18.20) | | | |
| | Judgments | 0.66 | 70.43 (16.85) | | | |
| | Total | | | | | |
| Property | Relevance | 0.80 | 65.53 (19.04) | | | |
| | Judgments | 0.53 | 61.22 (16.78) | | | |
| | Total | | | | | |

**Study 2**

| Foundation | Subscale | α | Total | Liberals | Moderates | Conservatives |
|---|---|---|---|---|---|---|
| Family | Relevance | 0.90 | 56.48 (26.07) | 52.43 (26.76) | 64.57 (21.61) | 60.11 (24.49) |
| | Judgments | 0.89 | 54.84 (25.71) | 49.9 (26.02) | 59.99 (22.74) | 63.47 (23.39) |
| | Total | 0.90 | 55.66 (23.41) | 51.16 (24) | 62.28 (19.41) | 61.79 (21.25) |
| Group | Relevance | 0.90 | 56.14 (25.36) | 57.89 (26.9) | 56.66 (22.38) | 53.11 (23.81) |
| | Judgments | 0.79 | 67.3 (19.03) | 67.02 (19.34) | 66.06 (19.34) | 69.99 (18.01) |
| | Total | 0.86 | 61.72 (19.55) | 62.45 (20.73) | 61.36 (18.81) | 61.55 (17.53) |
| Reciprocity | Relevance | 0.82 | 66.05 (21.17) | 65.48 (22.27) | 67.25 (19.07) | 66.7 (20.05) |
| | Judgments | 0.72 | 75.3 (15.91) | 75.69 (15.84) | 76 (16.37) | 73.67 (15.89) |
| | Total | 0.76 | 70.68 (15.15) | 70.58 (15.34) | 71.62 (15.25) | 70.18 (14.66) |
| Heroism | Relevance | 0.87 | 53.59 (25.01) | 51.38 (25.72) | 57.95 (21.11) | 55.52 (24.37) |
| | Judgments | 0.72 | 63.99 (21.26) | 59.69 (20.21) | 65.44 (21.76) | 73.76 (19.98) |
| | Total | 0.82 | 58.79 (19.84) | 55.53 (19.63) | 61.7 (18.79) | 64.64 (18.96) |
| Deference | Relevance | 0.84 | 41.74 (24.38) | 35.47 (23.84) | 44.25 (22.61) | 54.49 (22.31) |
| | Judgments | 0.77 | 44.89 (22.13) | 38.08 (20.1) | 46.19 (21.29) | 60.03 (19.17) |
| | Total | 0.87 | 43.32 (21.25) | 36.78 (19.81) | 45.22 (19.94) | 57.26 (18.39) |
| Fairness | Relevance | 0.77 | 56.01 (23.59) | 57.35 (23.7) | 58.22 (23.45) | 52.61 (22.85) |
| | Judgments | 0.77 | 80.97 (19.47) | 88.75 (14.31) | 77.79 (19.68) | 68.97 (20.35) |
| | Total | 0.72 | 68.49 (16.88) | 73.05 (14.9) | 68.01 (16.58) | 60.79 (17) |
| Property | Relevance | 0.83 | 69.49 (22.14) | 68.53 (23.5) | 72.12 (20.65) | 70.62 (18.11) |
| | Judgments | 0.79 | 48.61 (10.23) | 45.61 (9.19) | 51.27 (10.43) | 52.95 (10.07) |
| | Total | 0.72 | 59.05 (12.54) | 57.07 (13.04) | 61.69 (10.81) | 61.79 (11.05) |

*Note*s: Range for all subscales is 0–100. Standard deviations in parentheses. Cells are left empty where no data was available from the Curry et al (2019) paper

**Table 6.6.** *Correlations between the Relevance and Judgments subscales of MAC-Q*

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Curry et al (2019)** | | | | | | | |
| | MAC-Q Relevance subscales | | | | | | |
| | Family | Group | Reciprocity | Heroism | Deference | Fairness | Property |
| MAC-Q Judgments subscales | | | | | | | |
| Family | **0.49** | 0.24 | 0.27 | 0.29 | 0.35 | 0.14 | 0.19 |
| Group | 0.24 | **0.49** | 0.23 | 0.24 | 0.21 | 0.22 | 0.19 |
| Reciprocity | 0.41 | 0.37 | **0.34** | 0.34 | 0.24 | 0.22 | 0.25 |
| Heroism | 0.38 | 0.29 | 0.29 | **0.46** | 0.38 | 0.20 | 0.25 |
| Deference | 0.30 | 0.23 | 0.19 | 0.25 | **0.53** | 0.17 | 0.15 |
| Fairness | 0.22 | 0.23 | 0.18 | 0.14 | 0.07 | **0.15** | 0.10 |
| Property | - 0.01 | - 0.03 | 0.00 | - 0.01 | 0.02 | - 0.10 | **0.05** |
| After partialing ideology | | | | | | | |
| Family | | | | | | | |
| Group | | | | | | | |
| Reciprocity | | | | | | | |
| Heroism | | | | | | | |
| Deference | | | | | | | |
| Fairness | | | | | | | |
| Property | | | | | | | |
| **Study 2** | | | | | | | |
| | MAC-Q Relevance subscales | | | | | | |
| | Family | Group | Reciprocity | Heroism | Deference | Fairness | Property |
| MACQ Judgments subscales | | | | | | | |
| Family | **0.63** | 0.23 | 0.19 | 0.36 | 0.46 | 0.20 | 0.11 |
| Group | 0.34 | **0.54** | 0.27 | 0.39 | 0.23 | 0.34 | 0.20 |
| Reciprocity | 0.30 | 0.32 | 0.32 | 0.22 | 0.10 | 0.34 | **0.34** |
| Heroism | 0.41 | 0.26 | 0.26 | **0.47** | 0.43 | 0.17 | 0.18 |
| Deference | 0.35 | 0.09 | 0.12 | 0.28 | **0.67** | 0.12 | 0.06 |
| Fairness | 0.03 | 0.19 | 0.10 | 0.02 | -0.18 | **0.22** | 0.16 |
| Property | 0.03 | -0.06 | 0.06 | 0.01 | **0.26** | -0.12 | 0.08 |
| After partialing ideology | | | | | | | |
| Family | **0.64** | 0.24 | 0.19 | 0.36 | 0.45 | 0.22 | 0.10 |
| Group | 0.35 | **0.55** | 0.28 | 0.40 | 0.24 | 0.36 | 0.20 |
| Reciprocity | 0.31 | 0.32 | 0.33 | 0.23 | 0.11 | 0.34 | **0.34** |
| Heroism | 0.41 | 0.28 | 0.26 | **0.47** | 0.41 | 0.19 | 0.18 |
| Deference | 0.36 | 0.12 | 0.12 | 0.30 | **0.65** | 0.16 | 0.06 |
| Fairness | 0.06 | 0.16 | 0.12 | 0.04 | -0.10 | **0.20** | 0.19 |
| Property | 0.02 | -0.03 | 0.06 | 0.01 | **0.22** | -0.10 | 0.07 |

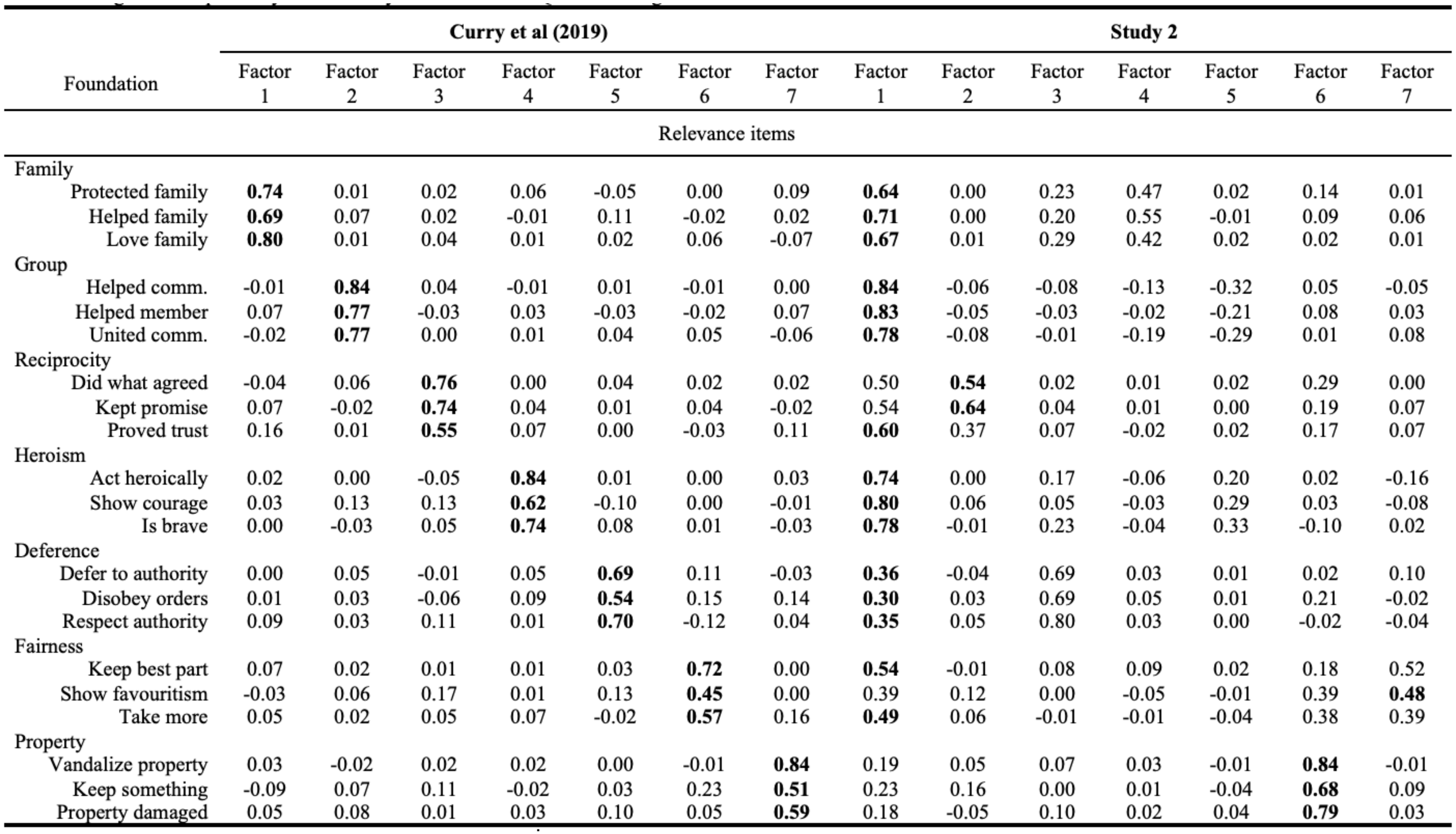
*Notes*: Each panel shows correlations between scales and partial correlations after controlling for political ideology. Highest correlation is shown in bold.

**Table 6.7.** *Factor loadings from Exploratory Factor Analysis for the MAC-Q items – Relevance items*

| | Curry et al (2019) | | | | | | | Study 2 | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Foundation | Factor 1 | Factor 2 | Factor 3 | Factor 4 | Factor 5 | Factor 6 | Factor 7 | Factor 1 | Factor 2 | Factor 3 | Factor 4 | Factor 5 | Factor 6 | Factor 7 |
| Judgment items | | | | | | | | | | | | | | |
| Family | | | | | | | | | | | | | | |
| Help family | **0.73** | 0.03 | 0.02 | -0.05 | 0.06 | 0.07 | -0.02 | **0.88** | 0.07 | 0.11 | 0.02 | -0.12 | 0.06 | -0.02 |
| Family loyalty | **0.80** | 0.01 | 0.02 | -0.02 | 0.00 | 0.00 | 0.07 | **0.90** | 0.08 | 0.05 | 0.05 | -0.07 | 0.01 | 0.02 |
| Family interest | **0.78** | -0.02 | 0.00 | 0.08 | -0.03 | -0.02 | -0.05 | **0.76** | -0.02 | 0.13 | 0.11 | -0.13 | -0.07 | -0.02 |
| Group | | | | | | | | | | | | | | |
| Help community | 0.04 | **0.71** | 0.00 | 0.00 | 0.04 | 0.04 | -0.09 | 0.31 | **0.68** | 0.23 | 0.05 | 0.06 | 0.16 | -0.14 |
| Com. participation | 0.00 | **0.79** | -0.05 | -0.01 | 0.02 | -0.03 | 0.01 | 0.20 | **0.81** | 0.29 | 0.14 | 0.03 | -0.01 | -0.05 |
| Useful member | -0.03 | **0.54** | 0.19 | 0.07 | -0.11 | 0.07 | 0.14 | 0.17 | **0.46** | 0.40 | 0.24 | -0.07 | 0.06 | 0.16 |
| Reciprocity | | | | | | | | | | | | | | |
| Reciprocate help | 0.16 | 0.14 | **0.42** | 0.05 | 0.05 | -0.05 | -0.13 | 0.23 | 0.24 | **0.55** | 0.11 | -0.02 | 0.11 | -0.07 |
| Make amends | -0.01 | 0.05 | **0.57** | 0.03 | 0.14 | 0.11 | 0.12 | 0.08 | 0.10 | **0.60** | 0.07 | 0.02 | 0.21 | 0.11 |
| Return favours | 0.09 | 0.02 | **0.70** | 0.04 | -0.02 | -0.03 | 0.01 | 0.12 | 0.11 | **0.79** | 0.10 | -0.01 | 0.17 | 0.08 |
| Heroism | | | | | | | | | | | | | | |
| Courage in adversity | 0.05 | 0.06 | 0.22 | **0.44** | -0.04 | 0.07 | -0.02 | 0.18 | 0.14 | 0.23 | **0.50** | -0.02 | 0.16 | -0.06 |
| Honour heroes | 0.05 | -0.04 | 0.09 | **0.68** | 0.01 | 0.03 | 0.02 | 0.42 | 0.07 | 0.18 | **0.64** | 0.02 | 0.02 | 0.01 |
| Sacrifice for country | -0.02 | 0.02 | -0.07 | **0.82** | 0.03 | -0.01 | -0.02 | 0.42 | 0.12 | 0.08 | **0.58** | 0.13 | -0.06 | 0.11 |
| Deference | | | | | | | | | | | | | | |
| Defer to superiors | -0.01 | 0.02 | 0.00 | -0.01 | **0.77** | 0.02 | -0.09 | **0.51** | 0.08 | 0.00 | 0.08 | 0.42 | -0.22 | -0.04 |
| Obedient to authority | 0.05 | -0.02 | 0.06 | 0.15 | **0.58** | -0.06 | 0.11 | **0.57** | 0.05 | -0.10 | 0.28 | 0.50 | -0.25 | 0.10 |
| Respect elders | 0.15 | 0.10 | 0.12 | 0.03 | **0.43** | 0.12 | 0.02 | **0.65** | 0.19 | 0.09 | 0.13 | 0.25 | 0.00 | 0.12 |
| Fairness | | | | | | | | | | | | | | |
| Equal treatment | 0.06 | -0.04 | -0.04 | 0.03 | 0.08 | **0.66** | -0.04 | 0.02 | 0.03 | 0.16 | 0.03 | 0.00 | **0.76** | 0.06 |
| Equal rights | 0.04 | 0.06 | -0.05 | 0.04 | -0.03 | **0.71** | 0.09 | -0.02 | -0.03 | 0.18 | 0.11 | -0.05 | **0.70** | 0.08 |
| Inequality unfair | -0.11 | 0.03 | 0.22 | -0.09 | -0.05 | **0.49** | -0.12 | -0.17 | 0.13 | 0.11 | -0.06 | -0.03 | **0.73** | -0.14 |
| Property | | | | | | | | | | | | | | |
| Steal food | 0.07 | 0.00 | -0.16 | 0.01 | 0.24 | -0.18 | **0.47** | 0.16 | 0.01 | 0.01 | 0.03 | 0.15 | -0.16 | **0.66** |
| Take things you need | 0.04 | -0.08 | 0.02 | -0.01 | -0.09 | 0.05 | **0.67** | 0.07 | 0.09 | -0.16 | 0.01 | 0.15 | -0.20 | **-0.65** |

*Notes:* Strongest factor loading for each item indicated in bold.

**Table 6.8.** *Factor loadings from Exploratory Factor Analysis for the MAC-Q items – Judgment items*



| Foundation | Curry et al (2019) | | | | | | | Study 2 | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Factor 1 | Factor 2 | Factor 3 | Factor 4 | Factor 5 | Factor 6 | Factor 7 | Factor 1 | Factor 2 | Factor 3 | Factor 4 | Factor 5 | Factor 6 | Factor 7 |
| | | | | | | Relevance items | | | | | | | | |
| Family | | | | | | | | | | | | | | |
| Protected family | **0.74** | 0.01 | 0.02 | 0.06 | -0.05 | 0.00 | 0.09 | **0.64** | 0.00 | 0.23 | 0.47 | 0.02 | 0.14 | 0.01 |
| Helped family | **0.69** | 0.07 | 0.02 | -0.01 | 0.11 | -0.02 | 0.02 | **0.71** | 0.00 | 0.20 | 0.55 | -0.01 | 0.09 | 0.06 |
| Love family | **0.80** | 0.01 | 0.04 | 0.01 | 0.02 | 0.06 | -0.07 | **0.67** | 0.01 | 0.29 | 0.42 | 0.02 | 0.02 | 0.01 |
| Group | | | | | | | | | | | | | | |
| Helped comm. | -0.01 | **0.84** | 0.04 | -0.01 | 0.01 | -0.01 | 0.00 | **0.84** | -0.06 | -0.08 | -0.13 | -0.32 | 0.05 | -0.05 |
| Helped member | 0.07 | **0.77** | -0.03 | 0.03 | -0.03 | -0.02 | 0.07 | **0.83** | -0.05 | -0.03 | -0.02 | -0.21 | 0.08 | 0.03 |
| United comm. | -0.02 | **0.77** | 0.00 | 0.01 | 0.04 | 0.05 | -0.06 | **0.78** | -0.08 | -0.01 | -0.19 | -0.29 | 0.01 | 0.08 |
| Reciprocity | | | | | | | | | | | | | | |
| Did what agreed | -0.04 | 0.06 | **0.76** | 0.00 | 0.04 | 0.02 | 0.02 | 0.50 | **0.54** | 0.02 | 0.01 | 0.02 | 0.29 | 0.00 |
| Kept promise | 0.07 | -0.02 | **0.74** | 0.04 | 0.01 | 0.04 | -0.02 | 0.54 | **0.64** | 0.04 | 0.01 | 0.00 | 0.19 | 0.07 |
| Proved trust | 0.16 | 0.01 | **0.55** | 0.07 | 0.00 | -0.03 | 0.11 | **0.60** | 0.37 | 0.07 | -0.02 | 0.02 | 0.17 | 0.07 |
| Heroism | | | | | | | | | | | | | | |
| Act heroically | 0.02 | 0.00 | -0.05 | **0.84** | 0.01 | 0.00 | 0.03 | **0.74** | 0.00 | 0.17 | -0.06 | 0.20 | 0.02 | -0.16 |
| Show courage | 0.03 | 0.13 | 0.13 | **0.62** | -0.10 | 0.00 | -0.01 | **0.80** | 0.06 | 0.05 | -0.03 | 0.29 | 0.03 | -0.08 |
| Is brave | 0.00 | -0.03 | 0.05 | **0.74** | 0.08 | 0.01 | -0.03 | **0.78** | -0.01 | 0.23 | -0.04 | 0.33 | -0.10 | 0.02 |
| Deference | | | | | | | | | | | | | | |
| Defer to authority | 0.00 | 0.05 | -0.01 | 0.05 | **0.69** | 0.11 | -0.03 | **0.36** | -0.04 | 0.69 | 0.03 | 0.01 | 0.02 | 0.10 |
| Disobey orders | 0.01 | 0.03 | -0.06 | 0.09 | **0.54** | 0.15 | 0.14 | **0.30** | 0.03 | 0.69 | 0.05 | 0.01 | 0.21 | -0.02 |
| Respect authority | 0.09 | 0.03 | 0.11 | 0.01 | **0.70** | -0.12 | 0.04 | **0.35** | 0.05 | 0.80 | 0.03 | 0.00 | -0.02 | -0.04 |
| Fairness | | | | | | | | | | | | | | |
| Keep best part | 0.07 | 0.02 | 0.01 | 0.01 | 0.03 | **0.72** | 0.00 | **0.54** | -0.01 | 0.08 | 0.09 | 0.02 | 0.18 | 0.52 |
| Show favouritism | -0.03 | 0.06 | 0.17 | 0.01 | 0.13 | **0.45** | 0.00 | 0.39 | 0.12 | 0.00 | -0.05 | -0.01 | 0.39 | **0.48** |
| Take more | 0.05 | 0.02 | 0.05 | 0.07 | -0.02 | **0.57** | 0.16 | **0.49** | 0.06 | -0.01 | -0.01 | -0.04 | 0.38 | 0.39 |
| Property | | | | | | | | | | | | | | |
| Vandalize property | 0.03 | -0.02 | 0.02 | 0.02 | 0.00 | -0.01 | **0.84** | 0.19 | 0.05 | 0.07 | 0.03 | -0.01 | **0.84** | -0.01 |
| Keep something | -0.09 | 0.07 | 0.11 | -0.02 | 0.03 | 0.23 | **0.51** | 0.23 | 0.16 | 0.00 | 0.01 | -0.04 | **0.68** | 0.09 |
| Property damaged | 0.05 | 0.08 | 0.01 | 0.03 | 0.10 | 0.05 | **0.59** | 0.18 | -0.05 | 0.10 | 0.02 | 0.04 | **0.79** | 0.03 |

*Notes:* Strongest factor loading for each item indicated in bold.

**Table 6.9.** *Goodness of Fit indices for alternative models of Confirmatory Factor Analyses for MAC-Q*

| **Full MAC-Q (all items)** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Model | $\chi^2$ | $df$ | RMSEA | AIC | BIC | CFI | TLI | SRMR |
| *Curry et al (2019) – Table S21 of Supplementary materials* | | | | | | | | |
| 1. Simple domains | 3,013.99 | 798 | 0.08 | 171,904.34 | 172,047.93 | 0.71 | | 0.10 |
| **2. Different but Related** | 1,352.11 | 728 | 0.04 | 169,892.12 | 170,104.08 | 0.92 | | 0.05 |
| *Study 2* | | | | | | | | |
| 1. Simple Domains | | 798 | | 145,483.00 | 146,064.89 | | | 0.11 |
| **2. Different but related** | | 728 | | 130,446.22 | 131,277.49 | | | 0.07 |

*Notes:* Model in bold is the best fitting one according to the indices reported in here.

## *A.3. Summary of MFT and MAC*

### *A.3.1. Moral Foundations Theory*

Haidt and Joseph (2004) developed Moral Foundations Theory (henceforth, MFT). MFT argues that people's morality is built upon several moral foundations, which are defined as a set of universal modules organised prior to experience, where each module processes information and outputs moral judgments in return[44]. This theory is a new synthesis of moral psychology considering four main discoveries made since the 1980's (for an extensive coverage, see Haidt, 2007 and 2013).

The first principle of MFT is that emotions and intuitions drive people's moral judgments, and that reasoning is a post-hoc process made to justify one's own moral views to others. Each moral foundation, operationalised by a module, represents the preparedness of individuals to acquire moral knowledge on a given topic, while the domain of such modules defines the relevant concepts for each moral foundation. The function of intuitions is to trigger emotions that activate modules, which, in return, output moral judgments.

The second tenet of MFT is that morality is wider in its scope than it was once considered. Whereas the classical literature focused only on topics regarding harm and fairness, MFT postulates five different foundations of moral systems: harm/care, fairness/reciprocity, ingroup/loyalty, authority/respect, and purity/chastity[45]. We briefly define the five foundations below by describing the domain of each module and the characteristic emotion that triggers each foundation. Each foundation is activated by an intuition, which triggers the characteristic emotion of that foundation. Each foundation, given its modular nature, processes the perceived situation and outputs a moral judgment when activated[46].

The domain of the *harm* foundation are situations where suffering or distress is being inflicted to (harm) or can be avoided by an action towards (care) someone. The emotion characteristic of this foundation is compassion. The domain of the *fairness* foundation are situations where cheating, cooperation and deception may arise in interactions with non-kin. Anger, guilt, and gratitude are the emotions linked to the

[44] Haidt and Bjorklund (2008) define a module as "*informationally encapsulated special purpose processing mechanisms*".

[45] For brevity, we use harm to refer to the harm/care foundation, fairness to refer to the fairness/reciprocity foundation, loyalty to refer to the ingroup/loyalty foundation, authority to refer to the authority/respect foundation and purity to refer to the purity/chastity foundation. The name of the foundations is the one reported in Haidt and Joseph (2006).

[46] For an in-depth coverage of the definition of each foundation, or on how foundations influence moral judgments, one can refer to Graham et al (2013).

fairness foundation. The *loyalty* foundation's domain is defined by situations that may suppose a threat to or may generate cohesion within a group. Emotions of group pride and rage towards traitors are characteristic for the loyalty foundation. The domain of the *authority* foundation is represented by situations in which hierarchy, dominance and submission are involved. Respect and fear are the emotions associated with this foundation. The domain of the *purity* foundation is defined by situations where pathogens can be avoided. Disgust is the emotion linked to the purity foundation.

The third principle of MFT is that morality glues people into communities. In that vein, MFT divides its five foundations into two groups: individualizing and binding foundations. The Individualizing foundations' aim is to protect individual rights, harm and fairness being the ones considered as individualizing: harm and unfairness are usually thought of as directed towards an individual. Loyalty, authority, and purity compose the binding foundations, and their purpose is to bind people into groups. MFT states that the latter group makes people reflect on their obligations to their societies and to stick to their role within a bigger entity.

The fourth tenet of MFT is that interpersonal relations shape people's moral judgments. When people interact, the moral judgments and moral reasoning of some can trigger new intuitions in others. These new intuitions may activate foundations that will help to shape an individual's moral judgments. Culture plays a prominent role in determining both people's intuitions and the importance of each moral foundation; as traditions, rituals and specific virtues are socialised throughout everyday interactions. Hence, people embedded in different cultures may cultivate a greater propensity to rely on different kinds of moral knowledge, which will help to explain different conceptions of morality in different societies.

### *A.3.2. Morality As cooperation theory*

MAC was built by accepting the core tenets of MFT: innateness of morality, primacy of intuitions over reason, operationalization of foundations as modules and a pluralistic view of morality. The main differences between MFT and MAC are the selection process of the foundations and the actual foundations themselves. Whereas foundations in MFT were chosen by selecting commonalities between previous theoretical accounts of morality, MAC foundations were chosen by identifying several cooperation problems in human social life and developing a module bespoke to each cooperation problem. As a result, MAC includes more foundations than MFT and each foundation is linked to regulating a specific cooperative issue. In contrast, MFT's regulation of behaviour is more limited in its scope, given that several cooperation problems identified in MAC do not clearly refer to a foundation of MFT.

MAC is composed of seven foundations. The seven foundations are based on cooperation problems discussed in game theory[47]. The *Family Values* foundation captures the cooperation problem of distributing resources between kin. More generally, problems concerning distribution of resources are captured in game theory by bargaining games such as ultimatum or dictator games. The foundation linked to such games is the *Fairness* foundation. The *Reciprocity* foundation is linked to social dilemma games, such as prisoner's dilemmas or public goods games. The *Group Loyalty* foundation is built upon coordination games and the *Heroism, Deference* and *Property Rights* foundations are linked to different aspects of problems of conflict resolution.

According to MAC, a cooperative action in any of the aforementioned games is to be seen as morally good and a selfish action as morally bad. A higher adherence to a given foundation shows a greater importance of that foundation for a given person, and, hence, moral judgments of cooperative and selfish actions on the relevant cooperation problems should be more extreme.

[47] Here we describe each foundation by pointing towards the cooperation problem related to each foundation. This is the equivalent of the "domain" of a foundation in MFT. For a deeper treatment of MAC's foundations, see Curry et al (2019).

### *A.4. Results regarding Moral Foundations not included in chapter 2*

#### *A.4.1. Moral Foundations Theory*

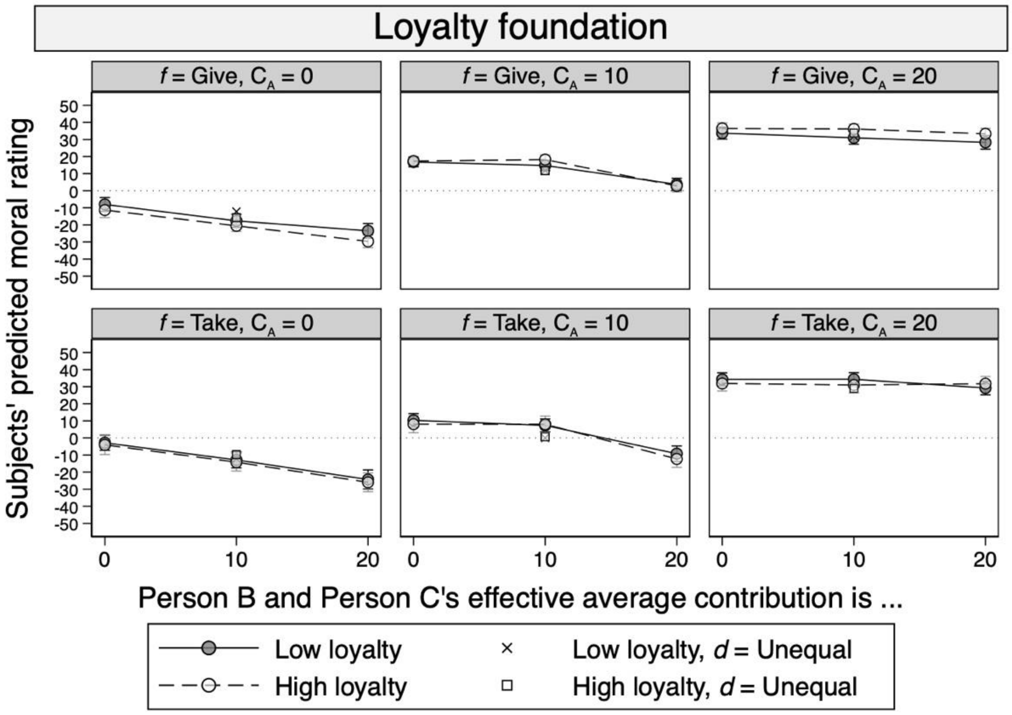


**Figure 6.1.** *Predicted MEF's evaluated at high and low scores of the Loyalty foundation – Top Panels: Give scenarios; Bottom Panels: Take scenarios*

Figure 6.1. reports the predicted moral judgments of high and low scores of the *Loyalty foundation*. We find only marginally significant effects when aggregating over all scenarios ($F(24{,}397) = 1.40$, $p = 0.10$), but no effects when considering Give and Take frames in isolation (Give frame: $F(12{,}397) = 1.44$, $p = 0.14$; Take frame: $F(12{,}397) = 1.35$; $p = 0.19$). Furthermore, we do not find any evidence of the Loyalty foundation being a moderator of the framing effects ($F(12{,}397) = 1.21$, $p = 0.27$). Overall, the effects of the Loyalty foundation in moral judgments are qualitatively similar to those of Harm and Fairness (i.e., higher scores in the Loyalty foundation leading to a harsher condemnation of free riders and a more pronounced praise of full contributors), although the size of the effect is smaller than that of Harm and Fairness.

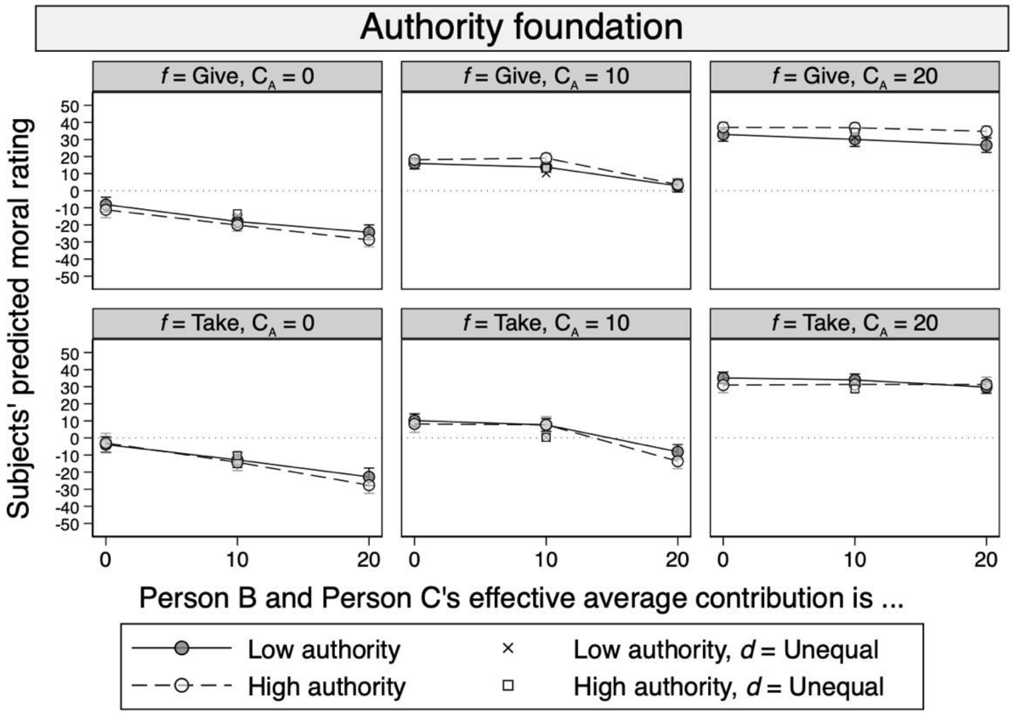


**Figure 6.2.** *Predicted MEF's evaluated at high and low scores of the Authority foundation – Top Panels: Give scenarios; Bottom Panels: Take scenarios*

Figure 6.2 reports the predicted moral judgments of high and low scores of the *Authority foundation*. We find evidence, both at the aggregate and at the frame level, of significant effects of the score of the Authority Foundation in our subjects' impartial moral judgments (Overall: $F(24{,}397) = 2.05$, $p = 0.00$; Give frame: $F(12{,}397) = 2.22$, $p = 0.01$; Take frame: $F(12{,}397) = 1.88$; $p = 0.03$). These results are mainly driven by the association between a higher score in the Authority foundation and (i) a higher praise of full contributors in the Give frame (see top right panel); (ii) a higher praise of half contributors when $C_{BC} = 10$ in the Give frame (see top central panel); and (iii) a higher condemnation of free riders and half contributors when $C_{BC} = 20$ in the Take frame (see bottom left and central panels). Furthermore, we do not find evidence in support of the Authority foundation as a moderator of framing effects ($F(12{,}397) = 1.08$; $p = 0.37$). However, there seems to be a tendency of a higher score (relative to a lower score) in the Authority foundation to accentuate the praise ascribed to full contributors in the Give scenarios.

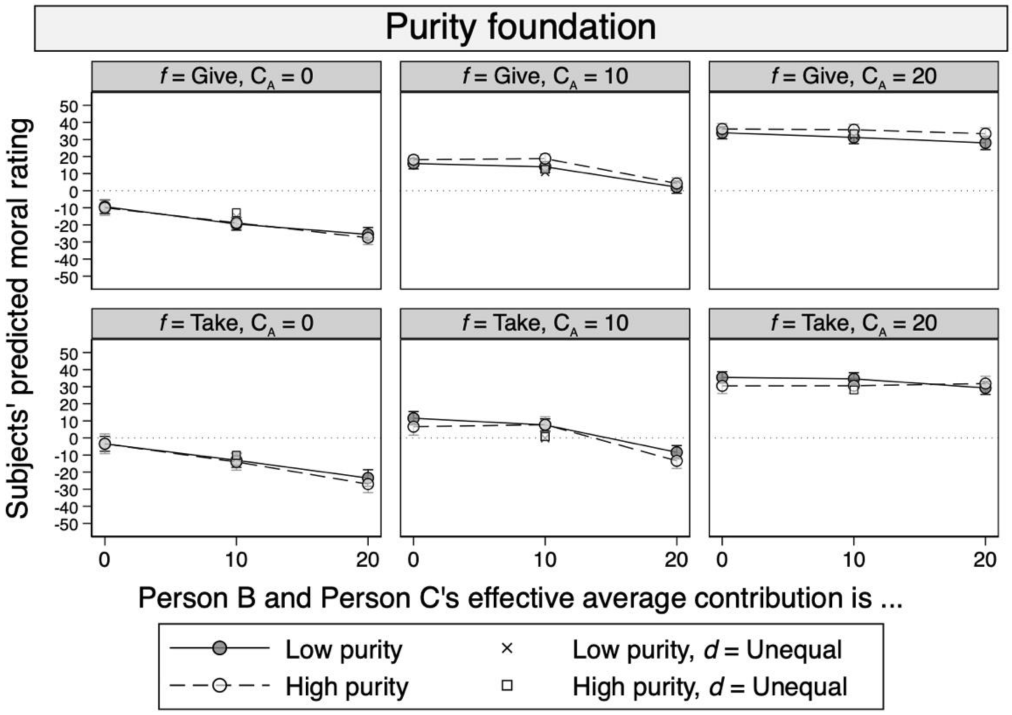


**Figure 6.3.** *Predicted MEF's evaluated at high and low scores of the Purity foundation – Top Panels: Give scenarios; Bottom Panels: Take scenarios*

Figure 6.3 reports the predicted moral judgments of high and low scores of the *Purity foundation.* We find overall evidence of Purity as an influence of moral judgments ($F(24{,}397) = 2.07$, $p = 0.00$), this mainly being driven by its effect on moral judgments in the Take frame (Give frame: $F(12{,}397) = 1.44$, $p = 0.14$; Take frame: $F(12{,}397) = 2.69$; $p = 0.00$). In this case, we also find evidence of Purity moderating framing effects ($F(12{,}397) = 2.04$; $p = 0.02$): relative to lower scores, higher scores in the Purity foundation accentuate the praise ascribed to half and full contributors in the Give scenarios.

Figure 6.4 reports the predicted moral judgments of high and low scores of the Economic liberty foundation. We find, overall, a marginally significant effect of the Economic liberty foundation in the moral judgments ($F(24{,}397) = 1.42$, $p = 0.09$), this being mainly driven by the moral judgments in the Give frame (Give frame: $F(12{,}397) = 2.10$, $p = 0.02$; Take frame: $F(12{,}397) = 0.74$; $p = 0.71$).

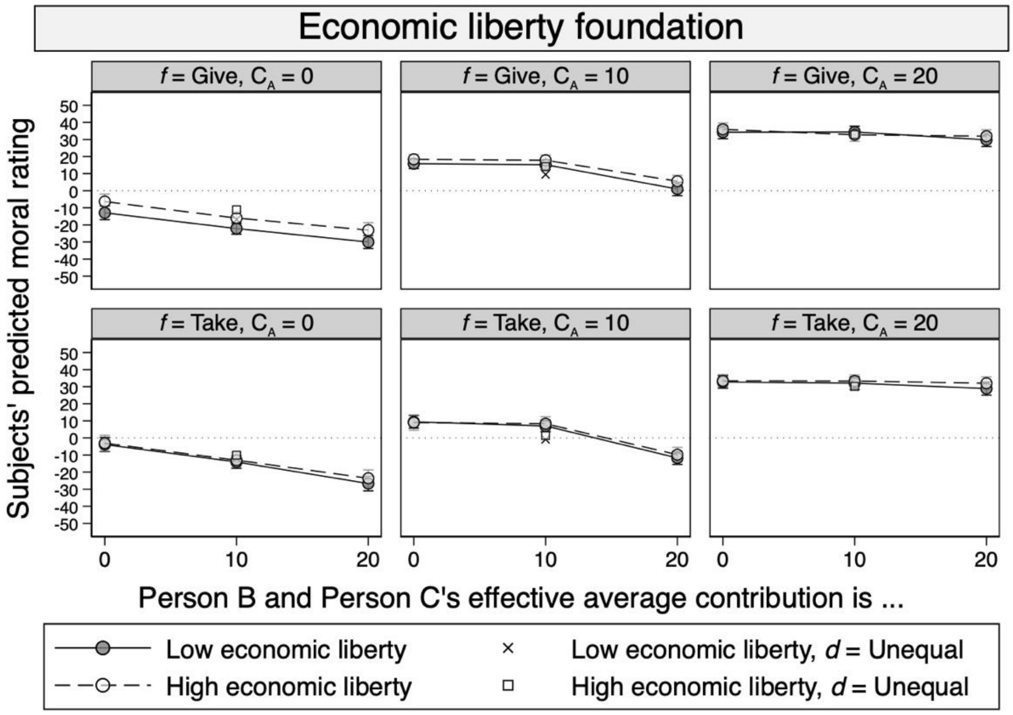


**Figure 6.4.** *Predicted MEF's evaluated at high and low scores of the Economic liberty foundation – Top Panels: Give scenarios; Bottom Panels: Take scenarios*

More specifically, the main effect of Economic liberty is channelled through its effect on Free riding in the Give frame: a higher score in the Economic liberty foundation was associated with a lower blame attached to free riders (see top left panel). We do not find evidence of the Economic liberty foundation moderating framing effects ($F(12{,}397) = 0.64$; $p = 0.82$). However, we notice that relative to a lower score, a higher score in the Economic liberty foundation decreases the framing effect of the MEF of free riding.

Finally, Figure 6.5 reports the predicted moral judgments of high and low scores of the *Lifestyle liberty foundation*. We did not find overall support for Lifestyle liberty as a driver of moral judgments of social dilemmas ($F(24{,}397) = 1.37$, $p = 0.12$), although we do find strong evidence of an effect of Lifestyle liberty in moral judgments of the Give frame (Give frame: $F(12{,}397) = 2.14$, $p = 0.01$; Take frame: $F(12{,}397) = 0.59$; $p = 0.85$). More specifically, we observe that a higher score in the Lifestyle liberty foundation increases the praise ascribed to full contributors in the Give frame (top right panel). We do not find any evidence is favour of the Lifestyle liberty foundation moderating framing effects ($F(12{,}397) = 0.69$; $p = 0.75$).

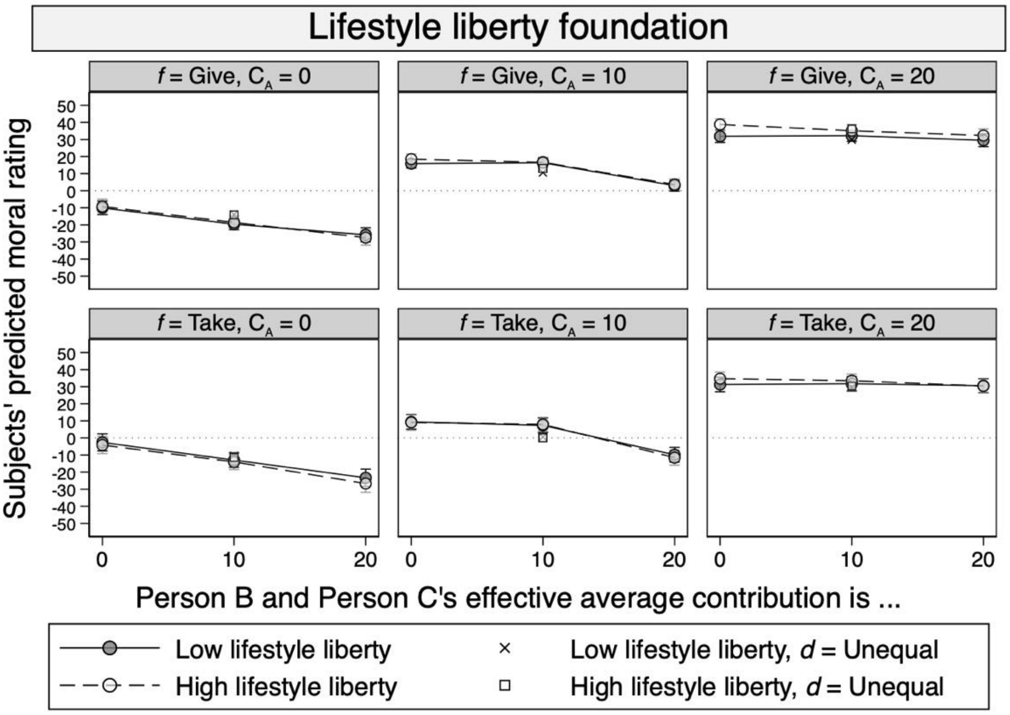


**Figure 6.5.** *Predicted MEF's evaluated at high and low scores of the Lifestyle liberty foundation – Top Panels: Give scenarios; Bottom Panels: Take scenarios*

*A.4.2. Morality As Cooperation Theory*

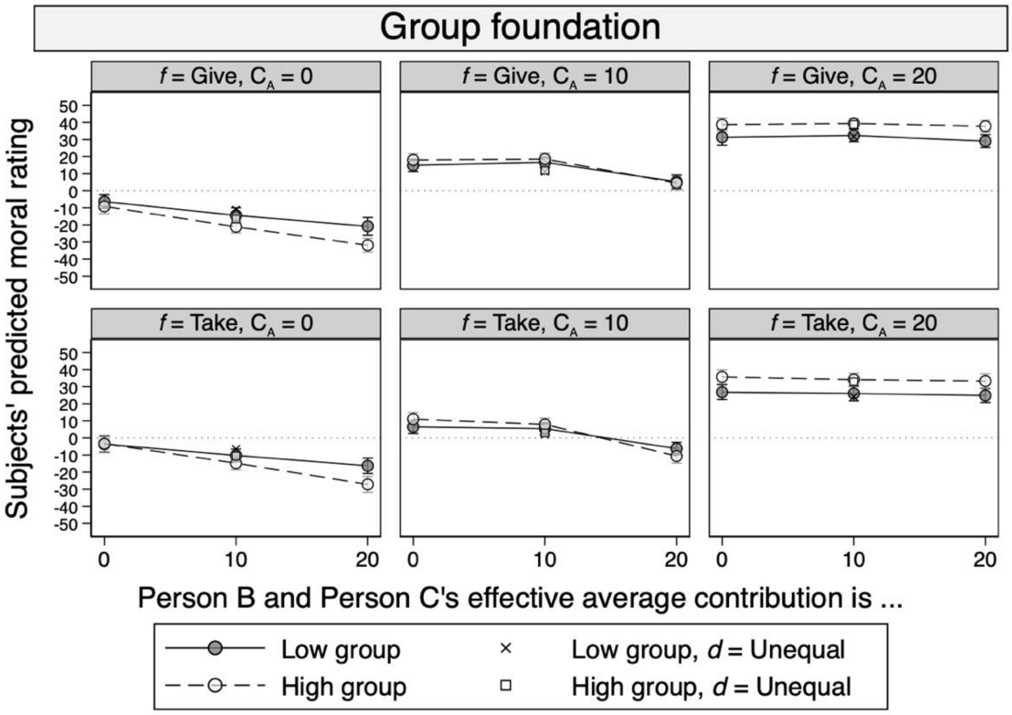


**Figure 6.6.** *Predicted MEF's evaluated at high and low scores of the Group foundation – Top Panels: Give scenarios; Bottom Panels: Take scenarios*

Figure 6.6. reports the predicted moral judgments of all scenarios of high and low levels of the *Group foundation*. As with the Fairness foundation, we do find evidence of a significant effect of Fairness on moral judgments; both at the aggregate level ($F(24{,}397) = 2.80,\ p = 0.00$), at both Give and Take scenarios (Give frame: $F(12{,}397) = 2.57, p = 0.00$; Take frame: $F(12{,}397) = 3.04$; $p = 0.00$), but we do not find evidence for the moderating role of the Group foundation in explaining framing effects ($F(12{,}397) = 0.32$; $p = 0.98$). As with the Fairness foundation, a higher score in the Group foundation is associated with (i) more blameworthy judgments of free riders in the Give and Take frames, in both cases driven by a more negative slope of the MEF of $C_A = 0$ (see top and bottom left panels); and (ii) more praiseworthy judgments of full contributors in both the Give and the Take frames (see top and bottom right panels). Overall, the effects of the Group foundation are smaller than the effects of either the Reciprocity or the Fairness foundation.

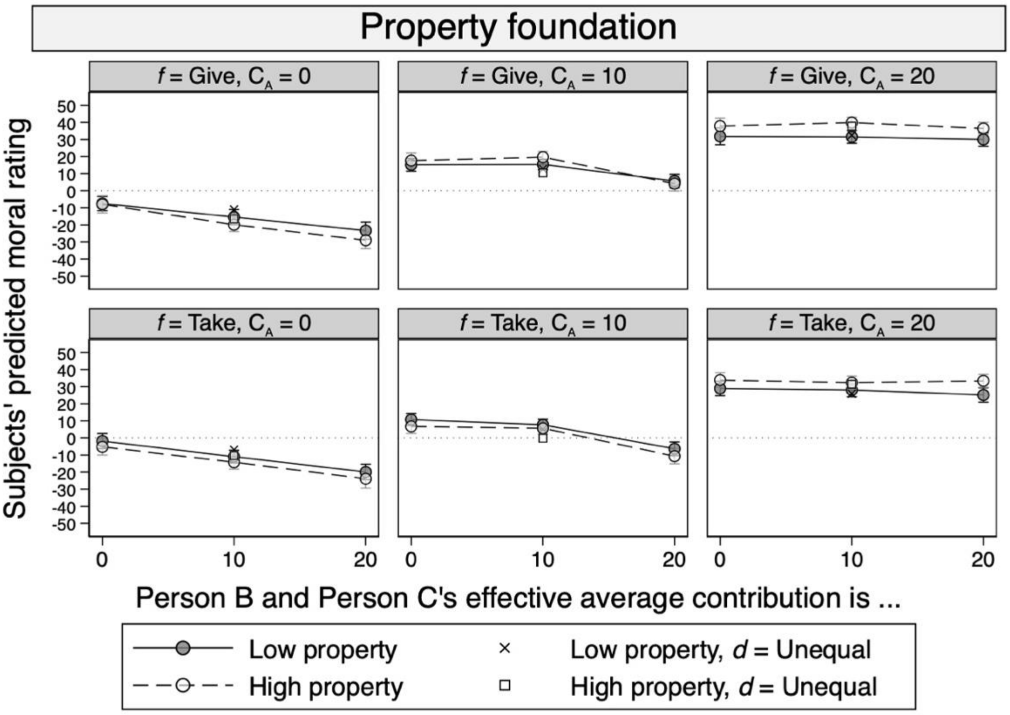


**Figure 6.7.** *Predicted MEF's evaluated at high and low scores of the Property foundation – Top Panels: Give scenarios; Bottom Panels: Take scenarios*

Figure 6.7. reports the predicted moral judgments of all scenarios of high and low levels of the *Property foundation*. As with the other foundations, we do find strong evidence of a significant effect of Fairness on moral judgments at the aggregate level ($F(24{,}397) = 1.77$, $p = 0.01$), but in this case the overall effect is mainly driven by its effect on the moral judgments of Take scenarios (Give frame: $F(12{,}397) = 1.48$, $p = 0.13$; Take frame: $F(12{,}397) = 2.05$; $p = 0.02$). The main visible effect of the Property foundation lies in the increase of praise ascribed to full contributors, especially in the Give frame (see top and bottom right panels). Furthermore, we do not find evidence supporting the moderating role of the Property foundation in framing effects ($F(12{,}397) = 1.10$; $p = 0.36$). Overall, the effects of the Property foundation are smaller than the effects of the previously discussed foundations.

Figure 6.8. reports the predicted moral judgments of all scenarios of high and low levels of the *Heroism foundation*. As with the other foundations, we do find evidence of a significant effect of Heroism on moral judgments at the aggregate level ($F(24{,}397) = 2.53$, $p = 0.00$), the overall effect being mainly driven by its effect on the moral judgments of Take scenarios, as it is the case with the Property foundation

(Give frame: $F(12{,}397) = 1.29$, $p = 0.22$; Take frame: $F(12{,}397) = 3.77$; $p = 0.00$).

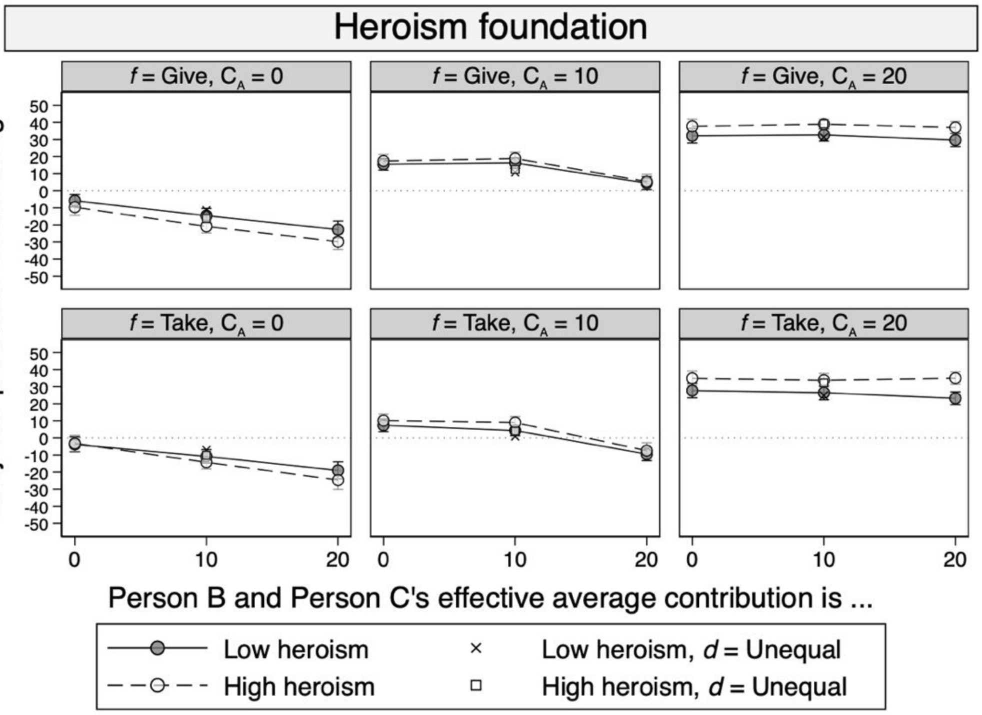


**Figure 6.8.** *Predicted MEF's evaluated at high and low scores of the Heroism foundation – Top Panels: Give scenarios; Bottom Panels: Take scenarios*

The main visible effect of the Heroism foundation lies in (i) the increasing blame of free riders, especially in the Give frame (see top and bottom left panels); and (ii) the increasing praise ascribed to full contributors, especially in the Take frame (see top and bottom right panels). Furthermore, we do not find evidence supporting the moderating role of the Heroism foundation in framing effects ($F(12{,}397) = 0.88$; $p = 0.57$). Overall, the effects of the Heroism foundation are similar in size to those of the Property foundation but smaller than the effects of the Reciprocity, Fairness, and Group foundations.

Figure 6.9. reports the predicted moral judgments of all scenarios of high and low levels of the *Deference foundation*. Contrasting with the findings of the other foundations, we do not find evidence of a significant effect of Fairness on moral judgments at the aggregate level ($F(24{,}397) = 1.38$, $p = 0.11$), and we only find an effect on the moral judgments of Give scenarios (Give frame: $F(12{,}397) = 1.74, p = 0.06$; Take frame: $F(12{,}397) = 1.02$; $p = 0.42$).

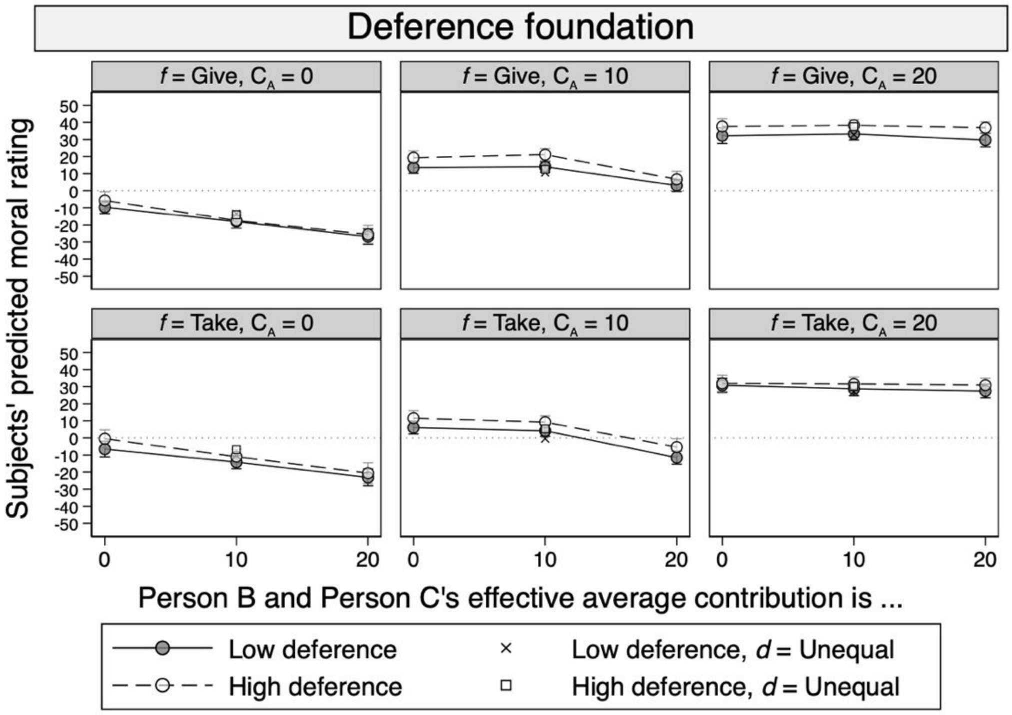


**Figure 6.9.** *Predicted MEF's evaluated at high and low scores of the Deference foundation – Top Panels: Give scenarios; Bottom Panels: Take scenarios*

The main visible effect of the Deference foundation lies in the increase of praise ascribed to half contributors in both frames, and of full contributors in the Give frame (see top and bottom central panels, and top right panel). Furthermore, we do not find evidence supporting a moderating role of the Deference foundation in framing effects ($F(12{,}397) = 0.58$; $p = 0.86$). Overall, the effects of the Deference foundation are similar to those of the Heroism and Property foundations, and smaller than the effects of the Reciprocity, Fairness, and Group foundations.

Finally, we discuss the role of high and low scores of the *Family foundation* in the moral judgments of the scenarios. Figure 6.10. reports those effects. In line with the majority of our findings, we find strong evidence of a significant effect of Family on moral judgments at the aggregate level ($F(24{,}397) = 1.84$, $p = 0.01$), this being driven by the effect of the Family foundation in Take scenarios (Give frame: $F(12{,}397) = 1.55$, $p = 0.10$; Take frame: $F(12{,}397) = 2.12$; $p = 0.01$).

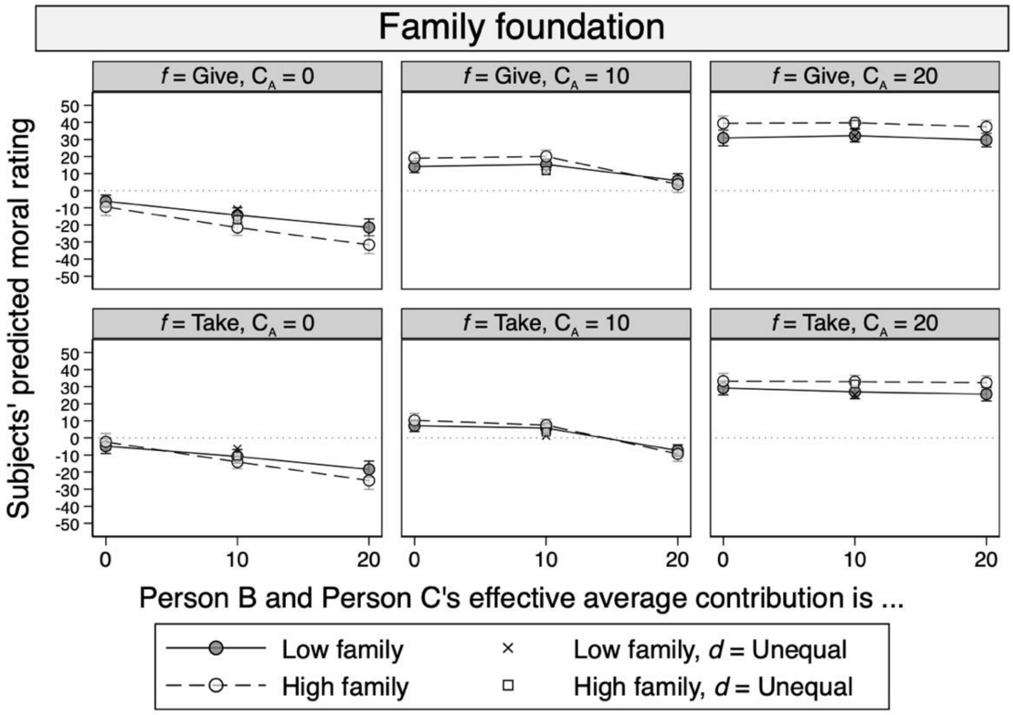


**Figure 6.10.** *Predicted MEF's evaluated at high and low scores of the Family foundation – Top Panels: Give scenarios; Bottom Panels: Take scenarios*

The main visible effect of the Family foundation lies in (i) the increase of blame ascribed to free riders, especially in the Give frame (see top and bottom left panels); (ii) the increase of praise of half contributors when $C_A \geq C_{Bc}$ in the Give frame (see top central panel); and (iii) the increase in praise of full contributors, especially prominent in the Give frame (see top and bottom right panels). Furthermore, we do not find evidence supporting a moderating role of the Family foundation in framing effects ($F(12{,}397) = 0.49$; $p = 0.92$). Overall, the effects of the Family foundation are similar to those of the Heroism, Deference and Property foundations, and smaller than the effects of the Reciprocity, Fairness, and Group foundations.

### *A.5. Rationale for tests of hypotheses regarding moral foundations*

#### *A.5.1. Formally testing for the effect of high versus low scores in moral foundations*

To test statistically whether high and low scores in a generic moral foundation $MF$ were statistically significantly different from 0 in a generic scenario $j = k$, we run the following statistical test:

$$H_0: m_{i,j=k}(z\_MF_i = -1) = m_{i,j=k}(z\_MF_i = +1)$$

Which, alternatively, can be rewritten as:

$$H_0: m_{i,j=k}(z\_MF_i = -1) - m_{i,j=k}(z\_MF_i = +1) = 0$$

By using the regression equation to substitute the two terms in the null hypothesis, we get:

$$H_0: \big(\beta_1 + \beta_k + \beta_{25} * (-1) + \beta_{24+k} * D_k * (-1)\big) - (\beta_1 + \beta_k + \beta_{25} + \beta_{24+k} * D_k) \\ = 0$$

Which, after simplifying, becomes:

$$H_0: (\beta_1 + \beta_k - \beta_{25} - \beta_{24+k} * D_k) - (\beta_1 + \beta_k + \beta_{25} + \beta_{24+k} * D_k) = 0$$

As $D_k$ is a dummy, and for scenario $j = k$ then $D_k = 1$, the previous equation further simplifies to:

$$H_0: (\beta_1 + \beta_k - \beta_{25} - \beta_{24+k}) - (\beta_1 + \beta_k + \beta_{25} + \beta_{24+k}) = 0$$

Expanding the parentheses, we get:

$$H_0: \beta_1 + \beta_k - \beta_{25} - \beta_{24+k} - \beta_1 - \beta_k - \beta_{25} - \beta_{24+k} = 0$$

Which, after cancelling out the relevant terms, becomes:

$$H_0: -\beta_{25} - \beta_{24+k} - \beta_{25} - \beta_{24+k} = 0$$

And, simplifying, we get:

$$H_0: -2\beta_{25} - 2\beta_{24+k} = 0$$

And, hence, the alternative hypothesis can, thus, be expressed as

$$H_1: -2\beta_{25} - 2\beta_{24+k} \neq 0$$

We performed this test for all scenarios. That is, for all $k \in [2,24]$.

And, for the baseline scenario ($k = 1$), that is for scenario $\langle C_A = 0, C_{BC} = 0, f = Give \rangle$, it follows from the same mathematical process highlighted above that the null and alternative hypotheses are:

$$H_0: -2\beta_{25} = 0$$

$$H_1: -2\beta_{25} \neq 0$$

For the joint test of hypothesis, we performed a test where the null hypothesis was that all the coefficients related to $z\_MF_i$ were jointly statistically insignificant from 0. That is,

$$H_0: \beta_l = 0 \; \forall \; l \in [25,48]$$

This yields an equivalent result to running a test where we impose that all the previous restrictions emanating from the individual statistical tests are jointly statistically insignificant from 0; as, in both cases, we are testing, either implicitly or explicitly, whether all the coefficients related to a generic moral foundation $MF$ are jointly significant or not.

Below we present tables regarding only all such tests for the individual hypotheses' tests of each scenario and each of the moral foundations, as the joint hypotheses tests have been presented in the main text. Given the multiple comparisons problem, we provide both the non-modified p-value and the Bonferroni-corrected p-value. We treat scenarios of different frames separately, hence Bonferroni corrected p-values consider 12 multiple comparisons of hypotheses tests at the same time (i.e., those of all the scenarios for a given frame).

**Table 6.10.** *Hypotheses tests regarding statistically different predicted moral judgments of each scenario for high and low scores in the Harm foundation of MFT – Study 1*

$H_0\colon m_{i,j=\langle C_A,C_{BC},f=Give,d\rangle}(z\_Harm_i = -1) = m_{i,j=\langle C_A,C_{BC},f=Give,d\rangle}(z\_Harm_i = +1)$

| Scenario | | | | Statistic | $p$-value | |
|---|---|---|---|---|---|---|
| $C_A$ | $C_{BC}$ | $f$ | $d$ | $F(1{,}397)$ | Normal | Bonferroni |
| 0 | 0 | Give | | 1.443 | 0.230 | 1 |
| 0 | 10 | Give | | 4.334 | 0.038** | 0.456 |
| 0 | 10 | Give | Unequal | 5.097 | 0.025** | 0.294 |
| 0 | 20 | Give | | 6.72 | 0.010** | 0.119 |
| 10 | 0 | Give | | 0.658 | 0.418 | 1 |
| 10 | 10 | Give | | 7.013 | 0.008*** | 0.101 |
| 10 | 10 | Give | Unequal | 0.002 | 0.961 | 1 |
| 10 | 20 | Give | | 0.031 | 0.861 | 1 |
| 20 | 0 | Give | | 6.378 | 0.012** | 0.143 |
| 20 | 10 | Give | | 7.193 | 0.008*** | 0.091* |
| 20 | 10 | Give | Unequal | 9.316 | 0.002*** | 0.029** |
| 20 | 20 | Give | | 8.766 | 0.003*** | 0.039** |

$H_0\colon m_{i,j=\langle C_A,C_{BC},f=Take,d\rangle}(z\_Harm_i = -1) = m_{i,j=\langle C_A,C_{BC},f=Take,d\rangle}(z\_Harm_i = +1)$

| Scenario | | | | Statistic | $p$-value | |
|---|---|---|---|---|---|---|
| $C_A$ | $C_{BC}$ | $f$ | $d$ | $F(1{,}397)$ | Normal | Bonferroni |
| 0 | 0 | Take | | 1.090 | 0.297 | 1 |
| 0 | 10 | Take | | 0.004 | 0.947 | 1 |
| 0 | 10 | Take | Unequal | 0.260 | 0.61 | 1 |
| 0 | 20 | Take | | 0.021 | 0.884 | 1 |
| 10 | 0 | Take | | 1.036 | 0.309 | 1 |
| 10 | 10 | Take | | 0.209 | 0.647 | 1 |
| 10 | 10 | Take | Unequal | 0.059 | 0.808 | 1 |
| 10 | 20 | Take | | 0.188 | 0.665 | 1 |
| 20 | 0 | Take | | 4.594 | 0.033** | 0.392 |
| 20 | 10 | Take | | 3.666 | 0.056* | 0.675 |
| 20 | 10 | Take | Unequal | 1.283 | 0.258 | 1 |
| 20 | 20 | Take | | 6.569 | 0.011** | 0.129 |

*Notes*: Statistical significance in the last two columns is defined by * $p$<0.1; ** $p$<0.05; *** $p$<0.01

**Table 6.11.** *Hypotheses tests regarding statistically different predicted moral judgments of each scenario for high and low scores in the Fairness foundation of MFT – Study 1*

$$H_0: m_{i,j=\langle C_A,C_{BC},f=Give,d\rangle}(z\_Fairness_i = -1) = m_{i,j=\langle C_A,C_{BC},f=Give,d\rangle}(z\_Fairness_i = +1)$$

| Scenario | | | | Statistic | $p$-value | |
|---|---|---|---|---|---|---|
| $C_A$ | $C_{BC}$ | $f$ | $d$ | $F(1{,}397)$ | Normal | Bonferroni |
| 0 | 0 | Give | | 5.190 | 0.023** | 0.279 |
| 0 | 10 | Give | | 2.997 | 0.084 | 1 |
| 0 | 10 | Give | Unequal | 6.609 | 0.011** | 0.126 |
| 0 | 20 | Give | | 5.082 | 0.025** | 0.297 |
| 10 | 0 | Give | | 0.406 | 0.524 | 1 |
| 10 | 10 | Give | | 1.491 | 0.223 | 1 |
| 10 | 10 | Give | Unequal | 0.022 | 0.883 | 1 |
| 10 | 20 | Give | | 0.03 | 0.863 | 1 |
| 20 | 0 | Give | | 4.888 | 0.028** | 0.331 |
| 20 | 10 | Give | | 6.189 | 0.013** | 0.159 |
| 20 | 10 | Give | Unequal | 5.867 | 0.016** | 0.190 |
| 20 | 20 | Give | | 1.364 | 0.244 | 1 |

$$H_0: m_{i,j=\langle C_A,C_{BC},f=Take,d\rangle}(z\_Fairness_i = -1) = m_{i,j=\langle C_A,C_{BC},f=Take,d\rangle}(z\_Fairness_i = +1)$$

| Scenario | | | | Statistic | $p$-value | |
|---|---|---|---|---|---|---|
| $C_A$ | $C_{BC}$ | $f$ | $d$ | $F(1{,}397)$ | Normal | Bonferroni |
| 0 | 0 | Take | | 0.337 | 0.562 | 1 |
| 0 | 10 | Take | | 0.646 | 0.422 | 1 |
| 0 | 10 | Take | Unequal | 0.685 | 0.408 | 1 |
| 0 | 20 | Take | | 0.787 | 0.375 | 1 |
| 10 | 0 | Take | | 0.005 | 0.943 | 1 |
| 10 | 10 | Take | | 0.021 | 0.884 | 1 |
| 10 | 10 | Take | Unequal | 1.307 | 0.254 | 1 |
| 10 | 20 | Take | | 0.282 | 0.595 | 1 |
| 20 | 0 | Take | | 2.39 | 0.123 | 1 |
| 20 | 10 | Take | | 4.544 | 0.034 | 0.404 |
| 20 | 10 | Take | Unequal | 2.868 | 0.091* | 1 |
| 20 | 20 | Take | | 7.443 | 0.007*** | 0.080* |

*Notes*: Statistical significance in the last two columns is defined by * $p$<0.1; ** $p$<0.05; *** $p$<0.01

**Table 6.12.** *Hypotheses tests regarding statistically different predicted moral judgments of each scenario for high and low scores in the Loyalty foundation of MFT – Study 1*

$$H_0: m_{i,j=\langle C_A,C_{BC},f=Give,d\rangle}(z\_Loyalty_i = -1) = m_{i,j=\langle C_A,C_{BC},f=Give,d\rangle}(z\_Loyalty_i = +1)$$

| Scenario | | | | Statistic | $p$-value | |
|---|---|---|---|---|---|---|
| $C_A$ | $C_{BC}$ | $f$ | $d$ | $F(1{,}397)$ | Normal | Bonferroni |
| 0 | 0 | Give | | 1.149 | 0.284 | 1 |
| 0 | 10 | Give | | 1.276 | 0.259 | 1 |
| 0 | 10 | Give | Unequal | 1.646 | 0.2 | 1 |
| 0 | 20 | Give | | 5.037 | 0.025** | 0.304 |
| 10 | 0 | Give | | 0.101 | 0.751 | 1 |
| 10 | 10 | Give | | 3.038 | 0.082* | 0.985 |
| 10 | 10 | Give | Unequal | 0.000 | 0.999 | 1 |
| 10 | 20 | Give | | 0.142 | 0.707 | 1 |
| 20 | 0 | Give | | 1.396 | 0.238 | 1 |
| 20 | 10 | Give | | 4.724 | 0.03** | 0.364 |
| 20 | 10 | Give | Unequal | 0.593 | 0.442 | 1 |
| 20 | 20 | Give | | 3.883 | 0.049** | 0.594 |

$$H_0: m_{i,j=\langle C_A,C_{BC},f=Take,d\rangle}(z\_Loyalty_i = -1) = m_{i,j=\langle C_A,C_{BC},f=Take,d\rangle}(z\_Loyalty_i = +1)$$

| Scenario | | | | Statistic | $p$-value | |
|---|---|---|---|---|---|---|
| $C_A$ | $C_{BC}$ | $f$ | $d$ | $F(1{,}397)$ | Normal | Bonferroni |
| 0 | 0 | Take | | 0.099 | 0.753 | 1 |
| 0 | 10 | Take | | 0.115 | 0.734 | 1 |
| 0 | 10 | Take | Unequal | 0.397 | 0.529 | 1 |
| 0 | 20 | Take | | 0.141 | 0.708 | 1 |
| 10 | 0 | Take | | 0.417 | 0.519 | 1 |
| 10 | 10 | Take | | 0.041 | 0.84 | 1 |
| 10 | 10 | Take | Unequal | 0.048 | 0.826 | 1 |
| 10 | 20 | Take | | 0.643 | 0.423 | 1 |
| 20 | 0 | Take | | 0.604 | 0.438 | 1 |
| 20 | 10 | Take | | 1.069 | 0.302 | 1 |
| 20 | 10 | Take | Unequal | 0.476 | 0.491 | 1 |
| 20 | 20 | Take | | 0.629 | 0.428 | 1 |

*Notes*: Statistical significance in the last two columns is defined by * $p$<0.1; ** $p$<0.05; *** $p$<0.01

**Table 6.13.** *Hypotheses tests regarding statistically different predicted moral judgments of each scenario for high and low scores in the Authority foundation of MFT – Study 1*

$$H_0: m_{i,j=\langle C_A,C_{BC},f=Give,d\rangle}(z\_Authority_i = -1) = m_{i,j=\langle C_A,C_{BC},f=Give,d\rangle}(z\_Authority_i = +1)$$

| Scenario | | | | Statistic | $p$-value | |
|---|---|---|---|---|---|---|
| $C_A$ | $C_{BC}$ | $f$ | $d$ | $F(1{,}397)$ | Normal | Bonferroni |
| 0 | 0 | Give | | 0.801 | 0.371 | 1 |
| 0 | 10 | Give | | 0.564 | 0.453 | 1 |
| 0 | 10 | Give | Unequal | 0.197 | 0.658 | 1 |
| 0 | 20 | Give | | 2.093 | 0.149 | 1 |
| 10 | 0 | Give | | 0.987 | 0.321 | 1 |
| 10 | 10 | Give | | 6.804 | 0.009*** | 0.113 |
| 10 | 10 | Give | Unequal | 1.862 | 0.173 | 1 |
| 10 | 20 | Give | | 0.076 | 0.783 | 1 |
| 20 | 0 | Give | | 2.686 | 0.102 | 1 |
| 20 | 10 | Give | | 6.829 | 0.009*** | 0.112 |
| 20 | 10 | Give | Unequal | 1.255 | 0.263 | 1 |
| 20 | 20 | Give | | 9.442 | 0.002*** | 0.027** |

$$H_0: m_{i,j=\langle C_A,C_{BC},f=Take,d\rangle}(z\_Authority_i = -1) = m_{i,j=\langle C_A,C_{BC},f=Take,d\rangle}(z\_Authority_i = +1)$$

| Scenario | | | | Statistic | $p$-value | |
|---|---|---|---|---|---|---|
| $C_A$ | $C_{BC}$ | $f$ | $d$ | $F(1{,}397)$ | Normal | Bonferroni |
| 0 | 0 | Take | | 0.054 | 0.816 | 1 |
| 0 | 10 | Take | | 0.154 | 0.695 | 1 |
| 0 | 10 | Take | Unequal | 0.122 | 0.727 | 1 |
| 0 | 20 | Take | | 1.726 | 0.19 | 1 |
| 10 | 0 | Take | | 0.364 | 0.547 | 1 |
| 10 | 10 | Take | | 0 | 0.985 | 1 |
| 10 | 10 | Take | Unequal | 0 | 1 | 1 |
| 10 | 20 | Take | | 2.838 | 0.093* | 1 |
| 20 | 0 | Take | | 2.229 | 0.136 | 1 |
| 20 | 10 | Take | | 0.888 | 0.347 | 1 |
| 20 | 10 | Take | Unequal | 0.577 | 0.448 | 1 |
| 20 | 20 | Take | | 0.271 | 0.603 | 1 |

*Notes*: Statistical significance in the last two columns is defined by * $p$<0.1; ** $p$<0.05; *** $p$<0.01

**Table 6.14.** *Hypotheses tests regarding statistically different predicted moral judgments of each scenario for high and low scores in the Purity foundation of MFT – Study 1*

$$H_0\colon m_{i,j=\langle C_A, C_{BC}, f=Give, d\rangle}(z\_Purity_i = -1) = m_{i,j=\langle C_A, C_{BC}, f=Give, d\rangle}(z\_Purity_i = +1)$$

| Scenario | | | | Statistic | $p$-value | |
|---|---|---|---|---|---|---|
| $C_A$ | $C_{BC}$ | $f$ | $d$ | $F(1{,}397)$ | Normal | Bonferroni |
| 0 | 0 | Give | | 0.040 | 0.841 | 1 |
| 0 | 10 | Give | | 0.078 | 0.780 | 1 |
| 0 | 10 | Give | Unequal | 0.779 | 0.378 | 1 |
| 0 | 20 | Give | | 0.493 | 0.483 | 1 |
| 10 | 0 | Give | | 1.186 | 0.277 | 1 |
| 10 | 10 | Give | | 5.121 | 0.024** | 0.290 |
| 10 | 10 | Give | Unequal | 0.616 | 0.433 | 1 |
| 10 | 20 | Give | | 0.772 | 0.380 | 1 |
| 20 | 0 | Give | | 0.931 | 0.335 | 1 |
| 20 | 10 | Give | | 3.296 | 0.070* | 0.843 |
| 20 | 10 | Give | Unequal | 0.175 | 0.676 | 1 |
| 20 | 20 | Give | | 4.164 | 0.042** | 0.504 |

$$H_0\colon m_{i,j=\langle C_A, C_{BC}, f=Take, d\rangle}(z\_Purity_i = -1) = m_{i,j=\langle C_A, C_{BC}, f=Take, d\rangle}(z\_Purity_i = +1)$$

| Scenario | | | | Statistic | $p$-value | |
|---|---|---|---|---|---|---|
| $C_A$ | $C_{BC}$ | $f$ | $d$ | $F(1{,}397)$ | Normal | Bonferroni |
| 0 | 0 | Take | | 0 | 0.994 | 1 |
| 0 | 10 | Take | | 0.084 | 0.772 | 1 |
| 0 | 10 | Take | Unequal | 0.169 | 0.681 | 1 |
| 0 | 20 | Take | | 0.956 | 0.329 | 1 |
| 10 | 0 | Take | | 2.197 | 0.139 | 1 |
| 10 | 10 | Take | | 0.003 | 0.954 | 1 |
| 10 | 10 | Take | Unequal | 0.093 | 0.760 | 1 |
| 10 | 20 | Take | | 2.595 | 0.108 | 1 |
| 20 | 0 | Take | | 2.994 | 0.084* | 1 |
| 20 | 10 | Take | | 1.762 | 0.185 | 1 |
| 20 | 10 | Take | Unequal | 1.088 | 0.298 | 1 |
| 20 | 20 | Take | | 0.722 | 0.396 | 1 |

*Notes*: Statistical significance in the last two columns is defined by * $p$<0.1; ** $p$<0.05; *** $p$<0.01

**Table 6.15.** *Hypotheses tests regarding statistically different predicted moral judgments of each scenario for high and low scores in the Economic liberty foundation of MFT – Study 1*

$$H_0\colon m_{i,j=\langle C_A,C_{BC},f=Give,d\rangle}(z\_Economic\ liberty_i = -1) = m_{i,j=\langle C_A,C_{BC},f=Give,d\rangle}(z\_Economic\ liberty_i = +1)$$

| Scenario | | | | Statistic | $p$-value | |
|---|---|---|---|---|---|---|
| $C_A$ | $C_{BC}$ | $f$ | $d$ | $F(1{,}397)$ | Normal | Bonferroni |
| 0 | 0 | Give | | 4.734 | 0.030** | 0.362 |
| 0 | 10 | Give | | 5.946 | 0.015** | 0.182 |
| 0 | 10 | Give | Unequal | 4.717 | 0.030** | 0.365 |
| 0 | 20 | Give | | 5.549 | 0.019** | 0.228 |
| 10 | 0 | Give | | 1.470 | 0.226 | 1 |
| 10 | 10 | Give | | 1.494 | 0.222 | 1 |
| 10 | 10 | Give | Unequal | 4.743 | 0.030** | 0.360 |
| 10 | 20 | Give | | 2.655 | 0.104 | 1 |
| 20 | 0 | Give | | 0.345 | 0.557 | 1 |
| 20 | 10 | Give | | 0.378 | 0.539 | 1 |
| 20 | 10 | Give | Unequal | 0 | 0.991 | 1 |
| 20 | 20 | Give | | 0.558 | 0.455 | 1 |

$$H_0\colon m_{i,j=\langle C_A,C_{BC},f=Take,d\rangle}(z\_Economic\ liberty_i = -1) = m_{i,j=\langle C_A,C_{BC},f=Take,d\rangle}(z\_Economic\ liberty_i = +1)$$

| Scenario | | | | Statistic | $p$-value | |
|---|---|---|---|---|---|---|
| $C_A$ | $C_{BC}$ | $f$ | $d$ | $F(1{,}397)$ | Normal | Bonferroni |
| 0 | 0 | Take | | 0.077 | 0.782 | 1 |
| 0 | 10 | Take | | 0.155 | 0.694 | 1 |
| 0 | 10 | Take | Unequal | 0.507 | 0.477 | 1 |
| 0 | 20 | Take | | 0.982 | 0.322 | 1 |
| 10 | 0 | Take | | 0.022 | 0.882 | 1 |
| 10 | 10 | Take | | 0.252 | 0.616 | 1 |
| 10 | 10 | Take | Unequal | 0.95 | 0.330 | 1 |
| 10 | 20 | Take | | 0.431 | 0.512 | 1 |
| 20 | 0 | Take | | 0.085 | 0.771 | 1 |
| 20 | 10 | Take | | 0.288 | 0.592 | 1 |
| 20 | 10 | Take | Unequal | 0.014 | 0.905 | 1 |
| 20 | 20 | Take | | 1.422 | 0.234 | 1 |

*Notes*: Statistical significance in the last two columns is defined by * *p*<0.1; ** *p*<0.05; *** *p*<0.01

**Table 6.16.** *Hypotheses tests regarding statistically different predicted moral judgments of each scenario for high and low scores in the Lifestyle liberty foundation of MFT – Study 1*

$H_0: m_{i,j=\langle C_A,C_{BC},f=Give,d\rangle}(z\_Lifestyle\ liberty_i = -1) = m_{i,j=\langle C_A,C_{BC},f=Give,d\rangle}(z\_Lifestyle\ liberty_i = +1)$

| Scenario | | | | Statistic | $p$-value | |
|---|---|---|---|---|---|---|
| $C_A$ | $C_{BC}$ | $f$ | $d$ | $F(1{,}397)$ | Normal | Bonferroni |
| 0 | 0 | Give | | 0.077 | 0.781 | 1 |
| 0 | 10 | Give | | 0.176 | 0.675 | 1 |
| 0 | 10 | Give | Unequal | 0.008 | 0.928 | 1 |
| 0 | 20 | Give | | 0.288 | 0.592 | 1 |
| 10 | 0 | Give | | 1.701 | 0.193 | 1 |
| 10 | 10 | Give | | 0.020 | 0.886 | 1 |
| 10 | 10 | Give | Unequal | 1.682 | 0.195 | 1 |
| 10 | 20 | Give | | 0.091 | 0.763 | 1 |
| 20 | 0 | Give | | 7.789 | 0.006*** | 0.066* |
| 20 | 10 | Give | | 1.440 | 0.231 | 1 |
| 20 | 10 | Give | Unequal | 8.021 | 0.005*** | 0.058* |
| 20 | 20 | Give | | 1.033 | 0.310 | 1 |

$H_0: m_{i,j=\langle C_A,C_{BC},f=Take,d\rangle}(z\_Lifestyle\ liberty_i = -1) = m_{i,j=\langle C_A,C_{BC},f=Take,d\rangle}(z\_Lifestyle\ liberty_i = +1)$

| Scenario | | | | Statistic | $p$-value | |
|---|---|---|---|---|---|---|
| $C_A$ | $C_{BC}$ | $f$ | $d$ | $F(1{,}397)$ | Normal | Bonferroni |
| 0 | 0 | Take | | 0.166 | 0.684 | 1 |
| 0 | 10 | Take | | 0.121 | 0.728 | 1 |
| 0 | 10 | Take | Unequal | 0.067 | 0.796 | 1 |
| 0 | 20 | Take | | 0.780 | 0.378 | 1 |
| 10 | 0 | Take | | 0.012 | 0.914 | 1 |
| 10 | 10 | Take | | 0.033 | 0.856 | 1 |
| 10 | 10 | Take | Unequal | 0.027 | 0.870 | 1 |
| 10 | 20 | Take | | 0.258 | 0.612 | 1 |
| 20 | 0 | Take | | 1.305 | 0.254 | 1 |
| 20 | 10 | Take | | 0.355 | 0.552 | 1 |
| 20 | 10 | Take | Unequal | 0.022 | 0.883 | 1 |
| 20 | 20 | Take | | 0.003 | 0.957 | 1 |

*Notes*: Statistical significance in the last two columns is defined by * $p$<0.1; ** $p$<0.05; *** $p$<0.01

**Table 6.17.** *Hypotheses tests regarding statistically different predicted moral judgments of each scenario for high and low scores in the Reciprocity foundation of MAC – Study 2*

$$H_0\colon m_{i,j=\langle C_A,C_{BC},f=Give,d\rangle}(z\_Reciprocity_i = -1) = m_{i,j=\langle C_A,C_{BC},f=Give,d\rangle}(z\_Reciprocity_i = +1)$$

| Scenario | | | | Statistic | $p$-value | |
|---|---|---|---|---|---|---|
| $C_A$ | $C_{BC}$ | $f$ | $d$ | $F(1{,}397)$ | Normal | Bonferroni |
| 0 | 0 | Give | | 0.193 | 0.661 | 1 |
| 0 | 10 | Give | | 6.103 | 0.014** | 0.167 |
| 0 | 10 | Give | Unequal | 3.337 | 0.068 | 0.822 |
| 0 | 20 | Give | | 12.631 | 0.000*** | 0.005*** |
| 10 | 0 | Give | | 3.394 | 0.066* | 0.794 |
| 10 | 10 | Give | | 4.638 | 0.032 | 0.383 |
| 10 | 10 | Give | Unequal | 0.555 | 0.457 | 1 |
| 10 | 20 | Give | | 0.525 | 0.469 | 1 |
| 20 | 0 | Give | | 6.543 | 0.011** | 0.131 |
| 20 | 10 | Give | | 21.175 | 0.000*** | 0.000*** |
| 20 | 10 | Give | Unequal | 18.014 | 0.000*** | 0.000*** |
| 20 | 20 | Give | | 8.241 | 0.004*** | 0.052* |

$$H_0\colon m_{i,j=\langle C_A,C_{BC},f=Take,d\rangle}(z\_Reciprocity_i = -1) = m_{i,j=\langle C_A,C_{BC},f=Take,d\rangle}(z\_Reciprocity_i = +1)$$

| Scenario | | | | Statistic | $p$-value | |
|---|---|---|---|---|---|---|
| $C_A$ | $C_{BC}$ | $f$ | $d$ | $F(1{,}397)$ | Normal | Bonferroni |
| 0 | 0 | Take | | 0.029 | 0.866 | 1 |
| 0 | 10 | Take | | 0.561 | 0.454 | 1 |
| 0 | 10 | Take | Unequal | 0.927 | 0.336 | 1 |
| 0 | 20 | Take | | 1.091 | 0.297 | 1 |
| 10 | 0 | Take | | 0.565 | 0.453 | 1 |
| 10 | 10 | Take | | 0.908 | 0.341 | 1 |
| 10 | 10 | Take | Unequal | 0.038 | 0.846 | 1 |
| 10 | 20 | Take | | 0.014 | 0.907 | 1 |
| 20 | 0 | Take | | 4.141 | 0.043** | 0.510 |
| 20 | 10 | Take | | 1.918 | 0.167 | 1 |
| 20 | 10 | Take | Unequal | 5.446 | 0.020** | 0.241 |
| 20 | 20 | Take | | 4.139 | 0.043** | 0.511 |

*Notes*: Statistical significance in the last two columns is defined by * $p$<0.1; ** $p$<0.05; *** $p$<0.01

**Table 6.18.** *Hypotheses tests regarding statistically different predicted moral judgments of each scenario for high and low scores in the Fairness foundation of MAC – Study 2*

$$H_0: m_{i,j=\langle C_A,C_{BC},f=Give,d\rangle}(z\_Fairness_i = -1) = m_{i,j=\langle C_A,C_{BC},f=Give,d\rangle}(z\_Fairness_i = +1)$$

| Scenario | | | | Statistic | $p$-value | |
|---|---|---|---|---|---|---|
| $C_A$ | $C_{BC}$ | $f$ | $d$ | $F(1{,}397)$ | Normal | Bonferroni |
| 0 | 0 | Give | | 1.295 | 0.256 | 1 |
| 0 | 10 | Give | | 21.541 | 0.000*** | 0.000*** |
| 0 | 10 | Give | Unequal | 16.402 | 0.000*** | 0.001*** |
| 0 | 20 | Give | | 35.355 | 0.000*** | 0.000*** |
| 10 | 0 | Give | | 2.024 | 0.156 | 1 |
| 10 | 10 | Give | | 2.345 | 0.127 | 1 |
| 10 | 10 | Give | Unequal | 1.054 | 0.305 | 1 |
| 10 | 20 | Give | | 4.088 | 0.044** | 0.526 |
| 20 | 0 | Give | | 12.367 | 0.000*** | 0.006*** |
| 20 | 10 | Give | | 19.601 | 0.000*** | 0.000*** |
| 20 | 10 | Give | Unequal | 14.435 | 0.000*** | 0.002*** |
| 20 | 20 | Give | | 9.17 | 0.003*** | 0.031** |

$$H_0: m_{i,j=\langle C_A,C_{BC},f=Take,d\rangle}(z\_Fairness_i = -1) = m_{i,j=\langle C_A,C_{BC},f=Take,d\rangle}(z\_Fairness_i = +1)$$

| Scenario | | | | Statistic | $p$-value | |
|---|---|---|---|---|---|---|
| $C_A$ | $C_{BC}$ | $f$ | $d$ | $F(1{,}397)$ | Normal | Bonferroni |
| 0 | 0 | Take | | 4.582 | 0.033** | 0.395 |
| 0 | 10 | Take | | 10.427 | 0.001*** | 0.016** |
| 0 | 10 | Take | Unequal | 7.648 | 0.006*** | 0.071* |
| 0 | 20 | Take | | 14.236 | 0.000*** | 0.002*** |
| 10 | 0 | Take | | 0.095 | 0.759 | 1 |
| 10 | 10 | Take | | 0.454 | 0.501 | 1 |
| 10 | 10 | Take | Unequal | 3.509 | 0.062 | 0.741 |
| 10 | 20 | Take | | 6.867 | 0.009*** | 0.109 |
| 20 | 0 | Take | | 17.926 | 0.000*** | 0.000*** |
| 20 | 10 | Take | | 18.904 | 0.000*** | 0.000*** |
| 20 | 10 | Take | Unequal | 32.821 | 0.000*** | 0.000*** |
| 20 | 20 | Take | | 16.638 | 0.000*** | 0.001*** |

*Notes*: Statistical significance in the last two columns is defined by * $p$<0.1; ** $p$<0.05; *** $p$<0.01

**Table 6.19.** *Hypotheses tests regarding statistically different predicted moral judgments of each scenario for high and low scores in the Group foundation of MAC – Study 2*

$$H_0: m_{i,j=\langle C_A,C_{BC},f=Give,d\rangle}(z\_Group_i = -1) = m_{i,j=\langle C_A,C_{BC},f=Give,d\rangle}(z\_Group_i = +1)$$

| Scenario | | | | Statistic | $p$-value | |
|---|---|---|---|---|---|---|
| $C_A$ | $C_{BC}$ | $f$ | $d$ | $F(1{,}397)$ | Normal | Bonferroni |
| 0 | 0 | Give | | 0.936 | 0.024** | 1 |
| 0 | 10 | Give | | 5.12 | 0.054* | 0.290 |
| 0 | 10 | Give | Unequal | 3.733 | 0.001*** | 0.649 |
| 0 | 20 | Give | | 10.978 | 0.246 | 0.012** |
| 10 | 0 | Give | | 1.349 | 0.388 | 1 |
| 10 | 10 | Give | | 0.746 | 0.747 | 1 |
| 10 | 10 | Give | Unequal | 0.104 | 0.794 | 1 |
| 10 | 20 | Give | | 0.068 | 0.023** | 1 |
| 20 | 0 | Give | | 5.19 | 0.003*** | 0.279 |
| 20 | 10 | Give | | 9.1 | 0.004*** | 0.033** |
| 20 | 10 | Give | Unequal | 8.583 | 0.001*** | 0.043** |
| 20 | 20 | Give | | 10.915 | 0.000*** | 0.012** |

$$H_0: m_{i,j=\langle C_A,C_{BC},f=Take,d\rangle}(z\_Group_i = -1) = m_{i,j=\langle C_A,C_{BC},f=Take,d\rangle}(z\_Group_i = +1)$$

| Scenario | | | | Statistic | $p$-value | |
|---|---|---|---|---|---|---|
| $C_A$ | $C_{BC}$ | $f$ | $d$ | $F(1{,}397)$ | Normal | Bonferroni |
| 0 | 0 | Take | | 0.008 | 0.927 | 1 |
| 0 | 10 | Take | | 3.129 | 0.078* | 0.932 |
| 0 | 10 | Take | Unequal | 1.839 | 0.176 | 1 |
| 0 | 20 | Take | | 13.002 | 0.000*** | 0.004*** |
| 10 | 0 | Take | | 2.649 | 0.104 | 1 |
| 10 | 10 | Take | | 0.865 | 0.353 | 1 |
| 10 | 10 | Take | Unequal | 0.119 | 0.731 | 1 |
| 10 | 20 | Take | | 2.944 | 0.087* | 1 |
| 20 | 0 | Take | | 8.733 | 0.003*** | 0.040** |
| 20 | 10 | Take | | 7.797 | 0.005*** | 0.066* |
| 20 | 10 | Take | Unequal | 10.46 | 0.001*** | 0.016** |
| 20 | 20 | Take | | 7.426 | 0.007*** | 0.081* |

*Notes*: Statistical significance in the last two columns is defined by * $p$<0.1; ** $p$<0.05; *** $p$<0.01

**Table 6.20.** *Hypotheses tests regarding statistically different predicted moral judgments of each scenario for high and low scores in the Property foundation of MAC – Study 2*

$$H_0\colon m_{i,j=\langle C_A,C_{BC},f=Give,d\rangle}(z\_Property_i = -1) = m_{i,j=\langle C_A,C_{BC},f=Give,d\rangle}(z\_Property_i = +1)$$

| Scenario | | | | Statistic | $p$-value | |
|---|---|---|---|---|---|---|
| $C_A$ | $C_{BC}$ | $f$ | $d$ | $F(1{,}397)$ | Normal | Bonferroni |
| 0 | 0 | Give | | 0.017 | 0.896 | 1 |
| 0 | 10 | Give | | 2.101 | 0.148 | 1 |
| 0 | 10 | Give | Unequal | 4.894 | 0.028** | 0.330 |
| 0 | 20 | Give | | 2.344 | 0.127 | 1 |
| 10 | 0 | Give | | 0.541 | 0.462 | 1 |
| 10 | 10 | Give | | 3.401 | 0.066* | 0.791 |
| 10 | 10 | Give | Unequal | 1.087 | 0.298 | 1 |
| 10 | 20 | Give | | 0.248 | 0.619 | 1 |
| 20 | 0 | Give | | 2.441 | 0.119 | 1 |
| 20 | 10 | Give | | 10.807 | 0.001*** | 0.013** |
| 20 | 10 | Give | Unequal | 5.131 | 0.024** | 0.288 |
| 20 | 20 | Give | | 5.469 | 0.020** | 0.238 |

$$H_0\colon m_{i,j=\langle C_A,C_{BC},f=Take,d\rangle}(z\_Property_i = -1) = m_{i,j=\langle C_A,C_{BC},f=Take,d\rangle}(z\_Property_i = +1)$$

| Scenario | | | | Statistic | $p$-value | |
|---|---|---|---|---|---|---|
| $C_A$ | $C_{BC}$ | $f$ | $d$ | $F(1{,}397)$ | Normal | Bonferroni |
| 0 | 0 | Take | | 1.099 | 0.295 | 1 |
| 0 | 10 | Take | | 1.398 | 0.238 | 1 |
| 0 | 10 | Take | Unequal | 1.466 | 0.227 | 1 |
| 0 | 20 | Take | | 1.217 | 0.271 | 1 |
| 10 | 0 | Take | | 2.111 | 0.147 | 1 |
| 10 | 10 | Take | | 0.766 | 0.382 | 1 |
| 10 | 10 | Take | Unequal | 5.27 | 0.022** | 0.267 |
| 10 | 20 | Take | | 2.046 | 0.153 | 1 |
| 20 | 0 | Take | | 2.479 | 0.116 | 1 |
| 20 | 10 | Take | | 2.305 | 0.130 | 1 |
| 20 | 10 | Take | Unequal | 4.474 | 0.035** | 0.420 |
| 20 | 20 | Take | | 7.833 | 0.005*** | 0.065* |

*Notes*: Statistical significance in the last two columns is defined by * $p$<0.1; ** $p$<0.05; *** $p$<0.01

**Table 6.21.** *Hypotheses tests regarding statistically different predicted moral judgments of each scenario for high and low scores in the Heroism foundation of MAC – Study 2*

$$H_0: m_{i,j=\langle C_A,C_{BC},f=Give,d\rangle}(z\_Heroism_i = -1) = m_{i,j=\langle C_A,C_{BC},f=Give,d\rangle}(z\_Heroism_i = +1)$$

| Scenario | | | | Statistic | $p$-value | |
|---|---|---|---|---|---|---|
| $C_A$ | $C_{BC}$ | $f$ | $d$ | $F(1{,}397)$ | Normal | Bonferroni |
| 0 | 0 | Give | | 1.614 | 0.205 | 1 |
| 0 | 10 | Give | | 4.552 | 0.033** | 0.402 |
| 0 | 10 | Give | Unequal | 3.403 | 0.066* | 0.790 |
| 0 | 20 | Give | | 4.182 | 0.042** | 0.498 |
| 10 | 0 | Give | | 0.521 | 0.471 | 1 |
| 10 | 10 | Give | | 1.16 | 0.282 | 1 |
| 10 | 10 | Give | Unequal | 0.679 | 0.411 | 1 |
| 10 | 20 | Give | | 0.085 | 0.771 | 1 |
| 20 | 0 | Give | | 3.041 | 0.082* | 0.983 |
| 20 | 10 | Give | | 7.16 | 0.008*** | 0.093* |
| 20 | 10 | Give | Unequal | 11.705 | 0.001*** | 0.008*** |
| 20 | 20 | Give | | 7.582 | 0.006*** | 0.074* |

$$H_0: m_{i,j=\langle C_A,C_{BC},f=Take,d\rangle}(z\_Heroism_i = -1) = m_{i,j=\langle C_A,C_{BC},f=Take,d\rangle}(z\_Heroism_i = +1)$$

| Scenario | | | | Statistic | $p$-value | |
|---|---|---|---|---|---|---|
| $C_A$ | $C_{BC}$ | $f$ | $d$ | $F(1{,}397)$ | Normal | Bonferroni |
| 0 | 0 | Take | | 0.034 | 0.854 | 1 |
| 0 | 10 | Take | | 1.593 | 0.208 | 1 |
| 0 | 10 | Take | Unequal | 1.232 | 0.268 | 1 |
| 0 | 20 | Take | | 2.04 | 0.154 | 1 |
| 10 | 0 | Take | | 1.02 | 0.313 | 1 |
| 10 | 10 | Take | | 4.394 | 0.037** | 0.440 |
| 10 | 10 | Take | Unequal | 2.427 | 0.120 | 1 |
| 10 | 20 | Take | | 0.466 | 0.495 | 1 |
| 20 | 0 | Take | | 5.644 | 0.018** | 0.216 |
| 20 | 10 | Take | | 5.937 | 0.015** | 0.183 |
| 20 | 10 | Take | Unequal | 7.28 | 0.007*** | 0.087* |
| 20 | 20 | Take | | 22.135 | 0.000*** | 0.000*** |

*Notes*: Statistical significance in the last two columns is defined by * $p$<0.1; ** $p$<0.05; *** $p$<0.01

**Table 6.22.** *Hypotheses tests regarding statistically different predicted moral judgments of each scenario for high and low scores in the Deference foundation of MAC – Study 2*

$$H_0\colon m_{i,j=\langle C_A, C_{BC}, f=Give, d\rangle}(z\_Deference_i = -1) = m_{i,j=\langle C_A, C_{BC}, f=Give, d\rangle}(z\_Deference_i = +1)$$

| Scenario | | | | Statistic | $p$-value | |
|---|---|---|---|---|---|---|
| $C_A$ | $C_{BC}$ | $f$ | $d$ | $F(1{,}397)$ | Normal | Bonferroni |
| 0 | 0 | Give | | 1.288 | 0.257 | 1 |
| 0 | 10 | Give | | 0.04 | 0.842 | 1 |
| 0 | 10 | Give | Unequal | 0.027 | 0.870 | 1 |
| 0 | 20 | Give | | 0.126 | 0.723 | 1 |
| 10 | 0 | Give | | 4.089 | 0.044** | 0.526 |
| 10 | 10 | Give | | 8.923 | 0.003*** | 0.036** |
| 10 | 10 | Give | Unequal | 0.67 | 0.414 | 1 |
| 10 | 20 | Give | | 1.394 | 0.238 | 1 |
| 20 | 0 | Give | | 2.31 | 0.129 | 1 |
| 20 | 10 | Give | | 4.187 | 0.041** | 0.497 |
| 20 | 10 | Give | Unequal | 3.825 | 0.051* | 0.614 |
| 20 | 20 | Give | | 7.109 | 0.008*** | 0.096* |

$$H_0\colon m_{i,j=\langle C_A, C_{BC}, f=Take, d\rangle}(z\_Deference_i = -1) = m_{i,j=\langle C_A, C_{BC}, f=Take, d\rangle}(z\_Deference_i = +1)$$

| Scenario | | | | Statistic | $p$-value | |
|---|---|---|---|---|---|---|
| $C_A$ | $C_{BC}$ | $f$ | $d$ | $F(1{,}397)$ | Normal | Bonferroni |
| 0 | 0 | Take | | 2.734 | 0.099* | 1 |
| 0 | 10 | Take | | 1.175 | 0.279 | 1 |
| 0 | 10 | Take | Unequal | 1.158 | 0.283 | 1 |
| 0 | 20 | Take | | 0.332 | 0.565 | 1 |
| 10 | 0 | Take | | 3.084 | 0.080* | 0.958 |
| 10 | 10 | Take | | 3.956 | 0.047** | 0.569 |
| 10 | 10 | Take | Unequal | 4.832 | 0.029** | 0.342 |
| 10 | 20 | Take | | 3.484 | 0.063* | 0.753 |
| 20 | 0 | Take | | 0.113 | 0.737 | 1 |
| 20 | 10 | Take | | 0.83 | 0.363 | 1 |
| 20 | 10 | Take | Unequal | 1.181 | 0.278 | 1 |
| 20 | 20 | Take | | 1.391 | 0.239 | 1 |

*Notes*: Statistical significance in the last two columns is defined by * $p$<0.1; ** $p$<0.05; *** $p$<0.01

**Table 6.23.** *Hypotheses tests regarding statistically different predicted moral judgments of each scenario for high and low scores in the Family foundation of MAC – Study 2*

$$H_0: m_{i,j=\langle C_A,C_{BC},f=Give,d\rangle}(z\_Family_i = -1) = m_{i,j=\langle C_A,C_{BC},f=Give,d\rangle}(z\_Family_i = +1)$$

| Scenario | | | | Statistic | $p$-value | |
|---|---|---|---|---|---|---|
| $C_A$ | $C_{BC}$ | $f$ | $d$ | $F(1{,}397)$ | Normal | Bonferroni |
| 0 | 0 | Give | | 0.99 | 0.320 | 1 |
| 0 | 10 | Give | | 4.399 | 0.037** | 0.439 |
| 0 | 10 | Give | Unequal | 3.615 | 0.058* | 0.696 |
| 0 | 20 | Give | | 6.845 | 0.009*** | 0.111 |
| 10 | 0 | Give | | 2.836 | 0.093* | 1 |
| 10 | 10 | Give | | 3.074 | 0.080* | 0.964 |
| 10 | 10 | Give | Unequal | 0.02 | 0.888 | 1 |
| 10 | 20 | Give | | 0.409 | 0.523 | 1 |
| 20 | 0 | Give | | 6.018 | 0.015** | 0.175 |
| 20 | 10 | Give | | 7.986 | 0.005*** | 0.059* |
| 20 | 10 | Give | Unequal | 6.27 | 0.013** | 0.152 |
| 20 | 20 | Give | | 6.412 | 0.012** | 0.141 |

$$H_0: m_{i,j=\langle C_A,C_{BC},f=Take,d\rangle}(z\_Family_i = -1) = m_{i,j=\langle C_A,C_{BC},f=Take,d\rangle}(z\_Family_i = +1)$$

| Scenario | | | | Statistic | $p$-value | |
|---|---|---|---|---|---|---|
| $C_A$ | $C_{BC}$ | $f$ | $d$ | $F(1{,}397)$ | Normal | Bonferroni |
| 0 | 0 | Take | | 0.572 | 0.450 | 1 |
| 0 | 10 | Take | | 1.32 | 0.251 | 1 |
| 0 | 10 | Take | Unequal | 2.117 | 0.146 | 1 |
| 0 | 20 | Take | | 3.278 | 0.071* | 0.851 |
| 10 | 0 | Take | | 1.441 | 0.231 | 1 |
| 10 | 10 | Take | | 0.558 | 0.456 | 1 |
| 10 | 10 | Take | Unequal | 1.015 | 0.314 | 1 |
| 10 | 20 | Take | | 0.537 | 0.464 | 1 |
| 20 | 0 | Take | | 1.64 | 0.201 | 1 |
| 20 | 10 | Take | | 4.489 | 0.035** | 0.417 |
| 20 | 10 | Take | Unequal | 5.551 | 0.019** | 0.228 |
| 20 | 20 | Take | | 5.445 | 0.020** | 0.241 |

*Notes*: Statistical significance in the last two columns is defined by * $p$<0.1; ** $p$<0.05; *** $p$<0.01

### *A.5.2. Formally testing for differential framing effects driven by moral foundations*

Additionally, we carry out some statistical tests to explore whether framing effects are different between low and high levels of a generic moral foundation $MF$, as this can provide some evidence on whether some moral foundations can be seen as an explanation of the framing effects we observe in the MEF's. Scenarios $k$ and $k + 12$ are scenarios that only vary on the framing of the decision situation. That is, scenarios that keep $C_A$, $C_{BC}$ and $d$ fixed while varying $f$. Taking this into account, we can then write the null hypothesis test concerning the equality of framing effects among high and low levels of a given moral foundation for scenarios $k$ and $k + 12$ as:

$$\begin{aligned} H_0\colon m_{i,k}(z\_MF_i = -1) - m_{i,k+12}(z\_MF_i = -1) \\ = m_{i,k}(z\_MF_i = +1) - m_{i,k+12}(z\_MF_i = +1) \end{aligned}$$

Where the left-hand side captures the framing effect for a low score in foundation $MF$ and the right-hand side captures the framing effect for a high score in foundation $MF$; the framing effect referring to the difference in moral judgments of scenarios $k$ and $k + 12$. As each frame has 12 moral judgments, we have 12 such hypothesis tests. Bringing all elements to the left-hand side, we get:

$$\begin{aligned} H_0\colon m_{i,k}(z\_MF_i = -1) - m_{i,k+12}(z\_MF_i = -1) - m_{i,k}(z\_MF_i = +1) \\ + m_{i,k+12}(z\_MF_i = +1) \end{aligned}$$

And, rearranging, the null hypothesis is, thus, equivalent to:

$$\begin{aligned} H_0\colon m_{i,k}(z\_MF_i = -1) - m_{i,k}(z\_MF_i = +1) \\ = m_{i,k+12}(z\_MF_i = -1) - m_{i,k+12}(z\_MF_i = +1) \end{aligned}$$

Hence, the null hypothesis test is equivalent to a hypothesis test where we impose that the differences between low and high scores in a given foundation are the same in scenarios $k$ (left-hand side) and $k + 12$ (right-hand side). Using, thus, the mathematical derivations outlined earlier, we can rewrite the null hypothesis as:

$$H_1\colon -2\beta_{25} - 2\beta_{24+k} = -2\beta_{25} - 2\beta_{24+k+12}$$

Adding the subscripts in the last element in the right-hand side, we get:

$$H_1: -2\beta_{25} - 2\beta_{24+k} = -2\beta_{25} - 2\beta_{36+k}$$

Bringing all elements to the left-hand side and cancelling out, we can rewrite the null and alternative hypotheses as:

$$H_0: -2\beta_{24+k} + 2\beta_{36+k} = 0$$

$$H_1: -2\beta_{24+k} + 2\beta_{36+k} \neq 0$$

Below we report the tables concerning the statistical tests regarding individual scenarios, as the joint tests of hypotheses for all foundations have been presented in the main text.

**Table 6.24.** *Hypotheses tests regarding statistically different predicted moral judgments of each scenario for high and low scores in the Harm foundation of MFT – Study 1*

$$H_0: m_{i,j=\langle C_A,C_{BC},f=Give,d\rangle}(z\_Harm_i = -1) - m_{i,j=\langle C_A,C_{BC},f=Take,d\rangle}(z\_Harm_i = -1)$$
$$= m_{i,j=\langle C_A,C_{BC},f=Give,d\rangle}(z\_Harm_i = +1) - m_{i,j=\langle C_A,C_{BC},f=Take,d\rangle}(z\_Harm_i = +1)$$

| Scenario | | | Statistic | $p$-value | |
|---|---|---|---|---|---|
| $C_A$ | $C_{BC}$ | | $F(1{,}397)$ | Normal | Bonferroni |
| 0 | 0 | | 2.439 | 0.119 | 1 |
| 0 | 10 | | 1.442 | 0.231 | 1 |
| 0 | 10 | Unequal | 0.856 | 0.355 | 1 |
| 0 | 20 | | 2.354 | 0.126 | 1 |
| 10 | 0 | | 0.128 | 0.721 | 1 |
| 10 | 10 | | 1.114 | 0.292 | 1 |
| 10 | 10 | Unequal | 0.030 | 0.863 | 1 |
| 10 | 20 | | 0.197 | 0.658 | 1 |
| 20 | 0 | | 0.042 | 0.837 | 1 |
| 20 | 10 | | 0.053 | 0.818 | 1 |
| 20 | 10 | Unequal | 0.901 | 0.343 | 1 |
| 20 | 20 | | 0.004 | 0.947 | 1 |

*Notes*: Statistical significance in the last two columns is defined by * $p$<0.1; ** $p$<0.05; *** $p$<0.01

**Table 6.25.** *Hypotheses tests regarding statistically different predicted moral judgments of each scenario for high and low scores in the Fairness foundation of MFT – Study 1*

$$H_0: m_{i,j=\langle C_A,C_{BC},f=Give,d\rangle}(z\_Fairness_i = -1) - m_{i,j=\langle C_A,C_{BC},f=Take,d\rangle}(z\_Fairness_i = -1)$$
$$= m_{i,j=\langle C_A,C_{BC},f=Give,d\rangle}(z\_Fairness_i = +1) - m_{i,j=\langle C_A,C_{BC},f=Take,d\rangle}(z\_Fairness_i = +1)$$

| Scenario | | | Statistic | $p$-value | |
|---|---|---|---|---|---|
| $C_A$ | $C_{BC}$ | | $F(1{,}397)$ | Normal | Bonferroni |
| 0 | 0 | | 2.905 | 0.089 | 1 |
| 0 | 10 | | 0.279 | 0.598 | 1 |
| 0 | 10 | Unequal | 0.931 | 0.335 | 1 |
| 0 | 20 | | 0.662 | 0.416 | 1 |
| 10 | 0 | | 0.166 | 0.684 | 1 |
| 10 | 10 | | 0.366 | 0.546 | 1 |
| 10 | 10 | Unequal | 0.671 | 0.413 | 1 |
| 10 | 20 | | 0.079 | 0.779 | 1 |
| 20 | 0 | | 0.223 | 0.637 | 1 |
| 20 | 10 | | 0.003 | 0.958 | 1 |
| 20 | 10 | Unequal | 0.005 | 0.944 | 1 |
| 20 | 20 | | 1.641 | 0.201 | 1 |

*Notes*: Statistical significance in the last two columns is defined by * *p*<0.1; ** *p*<0.05; *** *p*<0.01

**Table 6.26.** *Hypotheses tests regarding statistically different predicted moral judgments of each scenario for high and low scores in the Loyalty foundation of MFT – Study 1*

$$H_0: m_{i,j=\langle C_A,C_{BC},f=Give,d\rangle}(z\_Loyalty_i = -1) - m_{i,j=\langle C_A,C_{BC},f=Take,d\rangle}(z\_Loyalty_i = -1)$$
$$= m_{i,j=\langle C_A,C_{BC},f=Give,d\rangle}(z\_Loyalty_i = +1) - m_{i,j=\langle C_A,C_{BC},f=Take,d\rangle}(z\_Loyalty_i = +1)$$

| Scenario | | | Statistic | $p$-value | |
|---|---|---|---|---|---|
| $C_A$ | $C_{BC}$ | | $F(1{,}397)$ | Normal | Bonferroni |
| 0 | 0 | | 0.551 | 0.458 | 1 |
| 0 | 10 | | 0.115 | 0.735 | 1 |
| 0 | 10 | Unequal | 1.605 | 0.206 | 1 |
| 0 | 20 | | 0.764 | 0.383 | 1 |
| 10 | 0 | | 0.517 | 0.473 | 1 |
| 10 | 10 | | 0.525 | 0.469 | 1 |
| 10 | 10 | Unequal | 0.034 | 0.854 | 1 |
| 10 | 20 | | 0.214 | 0.644 | 1 |
| 20 | 0 | | 1.764 | 0.185 | 1 |
| 20 | 10 | | 4.490 | 0.035** | 0.417 |
| 20 | 10 | Unequal | 1.032 | 0.310 | 1 |
| 20 | 20 | | 0.397 | 0.529 | 1 |

*Notes*: Statistical significance in the last two columns is defined by * *p*<0.1; ** *p*<0.05; *** *p*<0.01

**Table 6.27.** *Hypotheses tests regarding statistically different predicted moral judgments of each scenario for high and low scores in the Authority foundation of MFT – Study 1*

$$H_0: m_{i,j=\langle C_A,C_{BC},f=Give,d\rangle}(z\_Authority_i = -1) - m_{i,j=\langle C_A,C_{BC},f=Take,d\rangle}(z\_Authority_i = -1)$$
$$= m_{i,j=\langle C_A,C_{BC},f=Give,d\rangle}(z\_Authority_i = +1) - m_{i,j=\langle C_A,C_{BC},f=Take,d\rangle}(z\_Authority_i = +1)$$

| Scenario | | | Statistic | $p$-value | |
|---|---|---|---|---|---|
| $C_A$ | $C_{BC}$ | | $F(1{,}397)$ | Normal | Bonferroni |
| 0 | 0 | | 0.794 | 0.374 | 1 |
| 0 | 10 | | 0.023 | 0.880 | 1 |
| 0 | 10 | Unequal | 0.000 | 0.991 | 1 |
| 0 | 20 | | 0.009 | 0.923 | 1 |
| 10 | 0 | | 1.102 | 0.295 | 1 |
| 10 | 10 | | 1.918 | 0.167 | 1 |
| 10 | 10 | Unequal | 0.598 | 0.440 | 1 |
| 10 | 20 | | 2.166 | 0.142 | 1 |
| 20 | 0 | | 4.873 | 0.028** | 0.334 |
| 20 | 10 | | 6.072 | 0.014** | 0.17 |
| 20 | 10 | Unequal | 1.611 | 0.205 | 1 |
| 20 | 20 | | 2.915 | 0.089* | 1 |

*Notes*: Statistical significance in the last two columns is defined by * *p*<0.1; ** *p*<0.05; *** *p*<0.01

**Table 6.28.** *Hypotheses tests regarding statistically different predicted moral judgments of each scenario for high and low scores in the Purity foundation of MFT – Study 1*

$$H_0: m_{i,j=\langle C_A,C_{BC},f=Give,d\rangle}(z\_Purity_i = -1) - m_{i,j=\langle C_A,C_{BC},f=Take,d\rangle}(z\_Purity_i = -1)$$
$$= m_{i,j=\langle C_A,C_{BC},f=Give,d\rangle}(z\_Purity_i = +1) - m_{i,j=\langle C_A,C_{BC},f=Take,d\rangle}(z\_Purity_i = +1)$$

| Scenario | | | Statistic | $p$-value | |
|---|---|---|---|---|---|
| $C_A$ | $C_{BC}$ | | $F(1{,}397)$ | Normal | Bonferroni |
| 0 | 0 | | 0.030 | 0.863 | 1 |
| 0 | 10 | | 0.159 | 0.690 | 1 |
| 0 | 10 | Unequal | 0.033 | 0.856 | 1 |
| 0 | 20 | | 0.119 | 0.730 | 1 |
| 10 | 0 | | 3.361 | 0.068 | 0.81 |
| 10 | 10 | | 1.594 | 0.207 | 1 |
| 10 | 10 | Unequal | 0.031 | 0.861 | 1 |
| 10 | 20 | | 3.271 | 0.071* | 0.855 |
| 20 | 0 | | 3.811 | 0.052* | 0.62 |
| 20 | 10 | | 4.734 | 0.030** | 0.362 |
| 20 | 10 | Unequal | 1.206 | 0.273 | 1 |
| 20 | 20 | | 0.515 | 0.473 | 1 |

*Notes*: Statistical significance in the last two columns is defined by * *p*<0.1; ** *p*<0.05; *** *p*<0.01

**Table 6.29.** *Hypotheses tests regarding statistically different predicted moral judgments of each scenario for high and low scores in the Economic liberty foundation of MFT – Study 1*

$$H_0: m_{i,j=\langle C_A,C_{BC},f=Give,d\rangle}(z\_Economic\ liberty_i = -1) - m_{i,j=\langle C_A,C_{BC},f=Take,d\rangle}(z\_Economic\ liberty_i = -1)$$
$$= m_{i,j=\langle C_A,C_{BC},f=Give,d\rangle}(z\_Economic\ liberty_i = +1) - m_{i,j=\langle C_A,C_{BC},f=Take,d\rangle}(z\_Economic\ liberty_i = +1)$$

| Scenario | | | Statistic | $p$-value | |
|---|---|---|---|---|---|
| $C_A$ | $C_{BC}$ | | $F(1{,}397)$ | Normal | Bonferroni |
| 0 | 0 | | 3.431 | 0.065 | 0.777 |
| 0 | 10 | | 1.941 | 0.164 | 1 |
| 0 | 10 | Unequal | 1.168 | 0.280 | 1 |
| 0 | 20 | | 0.817 | 0.367 | 1 |
| 10 | 0 | | 0.712 | 0.399 | 1 |
| 10 | 10 | | 0.164 | 0.686 | 1 |
| 10 | 10 | Unequal | 0.528 | 0.468 | 1 |
| 10 | 20 | | 0.443 | 0.506 | 1 |
| 20 | 0 | | 0.067 | 0.796 | 1 |
| 20 | 10 | | 0.665 | 0.415 | 1 |
| 20 | 10 | Unequal | 0.009 | 0.924 | 1 |
| 20 | 20 | | 0.063 | 0.802 | 1 |

*Notes*: Statistical significance in the last two columns is defined by * $p$<0.1; ** $p$<0.05; *** $p$<0.01

**Table 6.30.** *Hypotheses tests regarding statistically different predicted moral judgments of each scenario for high and low scores in the Lifestyle liberty foundation of MFT – Study 1*

$$H_0: m_{i,j=\langle C_A,C_{BC},f=Give,d\rangle}(z\_Lifestyle\ liberty_i = -1) - m_{i,j=\langle C_A,C_{BC},f=Take,d\rangle}(z\_Lifestyle\ liberty_i = -1)$$
$$= m_{i,j=\langle C_A,C_{BC},f=Give,d\rangle}(z\_Lifestyle\ liberty_i = +1) - m_{i,j=\langle C_A,C_{BC},f=Take,d\rangle}(z\_Lifestyle\ liberty_i = +1)$$

| Scenario | | | Statistic | $p$-value | |
|---|---|---|---|---|---|
| $C_A$ | $C_{BC}$ | | $F(1{,}397)$ | Normal | Bonferroni |
| 0 | 0 | | 0.207 | 0.649 | 1 |
| 0 | 10 | | 0.283 | 0.595 | 1 |
| 0 | 10 | Unequal | 0.022 | 0.882 | 1 |
| 0 | 20 | | 0.118 | 0.732 | 1 |
| 10 | 0 | | 0.593 | 0.442 | 1 |
| 10 | 10 | | 0.007 | 0.933 | 1 |
| 10 | 10 | Unequal | 0.704 | 0.402 | 1 |
| 10 | 20 | | 0.348 | 0.555 | 1 |
| 20 | 0 | | 0.829 | 0.363 | 1 |
| 20 | 10 | | 0.091 | 0.763 | 1 |
| 20 | 10 | Unequal | 2.226 | 0.136 | 1 |
| 20 | 20 | | 0.537 | 0.464 | 1 |

*Notes*: Statistical significance in the last two columns is defined by * $p$<0.1; ** $p$<0.05; *** $p$<0.01

**Table 6.31.** *Hypotheses tests regarding statistically different predicted moral judgments of each scenario for high and low scores in the Reciprocity foundation of MAC – Study 2*

$$H_0: m_{i,j=\langle C_A,C_{BC},f=Give,d\rangle}(z\_Reciprocity_i = -1) - m_{i,j=\langle C_A,C_{BC},f=Take,d\rangle}(z\_Reciprocity_i = -1)$$
$$= m_{i,j=\langle C_A,C_{BC},f=Give,d\rangle}(z\_Reciprocity_i = +1) - m_{i,j=\langle C_A,C_{BC},f=Take,d\rangle}(z\_Reciprocity_i = +1)$$

| Scenario | | | Statistic | $p$-value | |
|---|---|---|---|---|---|
| $C_A$ | $C_{BC}$ | | $F(1{,}397)$ | Normal | Bonferroni |
| 0 | 0 | | 0.112 | 0.738 | 1 |
| 0 | 10 | | 1.981 | 0.160 | 1 |
| 0 | 10 | Unequal | 0.227 | 0.634 | 1 |
| 0 | 20 | | 2.895 | 0.090* | 1 |
| 10 | 0 | | 0.852 | 0.356 | 1 |
| 10 | 10 | | 0.606 | 0.437 | 1 |
| 10 | 10 | Unequal | 0.431 | 0.512 | 1 |
| 10 | 20 | | 0.224 | 0.636 | 1 |
| 20 | 0 | | 0.536 | 0.465 | 1 |
| 20 | 10 | | 2.738 | 0.099 | 1 |
| 20 | 10 | Unequal | 0.862 | 0.354 | 1 |
| 20 | 20 | | 0.263 | 0.609 | 1 |

*Notes*: Statistical significance in the last two columns is defined by * *p*<0.1; ** *p*<0.05; *** *p*<0.01

***Table 6.32.*** *Hypotheses tests regarding statistically different predicted moral judgments of each scenario for high and low scores in the Fairness foundation of MAC – Study 2*

$$H_0: m_{i,j=\langle C_A,C_{BC},f=Give,d\rangle}(z\_Fairness_i = -1) - m_{i,j=\langle C_A,C_{BC},f=Take,d\rangle}(z\_Fairness_i = -1)$$
$$= m_{i,j=\langle C_A,C_{BC},f=Give,d\rangle}(z\_Fairness_i = +1) - m_{i,j=\langle C_A,C_{BC},f=Take,d\rangle}(z\_Fairness_i = +1)$$

| Scenario | | | Statistic | $p$-value | |
|---|---|---|---|---|---|
| $C_A$ | $C_{BC}$ | | $F(1{,}397)$ | Normal | Bonferroni |
| 0 | 0 | | 0.000 | 0.996 | 1 |
| 0 | 10 | | 0.794 | 0.373 | 1 |
| 0 | 10 | Unequal | 0.128 | 0.721 | 1 |
| 0 | 20 | | 1.123 | 0.290 | 1 |
| 10 | 0 | | 1.517 | 0.219 | 1 |
| 10 | 10 | | 2.360 | 0.125 | 1 |
| 10 | 10 | Unequal | 0.421 | 0.517 | 1 |
| 10 | 20 | | 0.142 | 0.706 | 1 |
| 20 | 0 | | 0.013 | 0.909 | 1 |
| 20 | 10 | | 0.048 | 0.827 | 1 |
| 20 | 10 | Unequal | 1.577 | 0.210 | 1 |
| 20 | 20 | | 0.338 | 0.562 | 1 |

*Notes*: Statistical significance in the last two columns is defined by * *p*<0.1; ** *p*<0.05; *** *p*<0.01

**Table 6.33.** *Hypotheses tests regarding statistically different predicted moral judgments of each scenario for high and low scores in the Group foundation of MAC – Study 2*

$$H_0\colon m_{i,j=\langle C_A,C_{BC},f=Give,d\rangle}(z\_Group_i=-1)-m_{i,j=\langle C_A,C_{BC},f=Take,d\rangle}(z\_Group_i=-1)$$
$$=m_{i,j=\langle C_A,C_{BC},f=Give,d\rangle}(z\_Group_i=+1)-m_{i,j=\langle C_A,C_{BC},f=Take,d\rangle}(z\_Group_i=+1)$$

| Scenario | | | Statistic | $p$-value | |
|---|---|---|---|---|---|
| $C_A$ | $C_{BC}$ | | $F(1{,}397)$ | Normal | Bonferroni |
| 0 | 0 | | 0.773 | 0.380 | 1 |
| 0 | 10 | | 0.297 | 0.586 | 1 |
| 0 | 10 | Unequal | 0.092 | 0.761 | 1 |
| 0 | 20 | | 0.001 | 0.970 | 1 |
| 10 | 0 | | 0.134 | 0.714 | 1 |
| 10 | 10 | | 0.025 | 0.875 | 1 |
| 10 | 10 | Unequal | 0.004 | 0.952 | 1 |
| 10 | 20 | | 0.872 | 0.351 | 1 |
| 20 | 0 | | 0.124 | 0.725 | 1 |
| 20 | 10 | | 0.095 | 0.758 | 1 |
| 20 | 10 | Unequal | 0.245 | 0.621 | 1 |
| 20 | 20 | | 0.004 | 0.950 | 1 |

*Notes*: Statistical significance in the last two columns is defined by * *p*<0.1; ** *p*<0.05; *** *p*<0.01

**Table 6.34.** *Hypotheses tests regarding statistically different predicted moral judgments of each scenario for high and low scores in the Property foundation of MAC – Study 2*

$$H_0\colon m_{i,j=\langle C_A,C_{BC},f=Give,d\rangle}(z\_Property_i=-1)-m_{i,j=\langle C_A,C_{BC},f=Take,d\rangle}(z\_Property_i=-1)$$
$$=m_{i,j=\langle C_A,C_{BC},f=Give,d\rangle}(z\_Property_i=+1)-m_{i,j=\langle C_A,C_{BC},f=Take,d\rangle}(z\_Property_i=+1)$$

| Scenario | | | Statistic | $p$-value | |
|---|---|---|---|---|---|
| $C_A$ | $C_{BC}$ | | $F(1{,}397)$ | Normal | Bonferroni |
| 0 | 0 | | 0.103 | 0.748 | 1 |
| 0 | 10 | | 0.093 | 0.761 | 1 |
| 0 | 10 | Unequal | 0.323 | 0.570 | 1 |
| 0 | 20 | | 0.102 | 0.749 | 1 |
| 10 | 0 | | 2.241 | 0.135 | 1 |
| 10 | 10 | | 3.676 | 0.056 | 0.671 |
| 10 | 10 | Unequal | 0.882 | 0.348 | 1 |
| 10 | 20 | | 0.447 | 0.504 | 1 |
| 20 | 0 | | 0.054 | 0.817 | 1 |
| 20 | 10 | | 1.148 | 0.285 | 1 |
| 20 | 10 | Unequal | 0.002 | 0.963 | 1 |
| 20 | 20 | | 0.193 | 0.660 | 1 |

*Notes*: Statistical significance in the last two columns is defined by * *p*<0.1; ** *p*<0.05; *** *p*<0.01

**Table 6.35.** *Hypotheses tests regarding statistically different predicted moral judgments of each scenario for high and low scores in the Heroism foundation of MAC – Study 2*

$$H_0\colon m_{i,j=\langle C_A,C_{BC},f=Give,d\rangle}(z\_Heroism_i = -1) - m_{i,j=\langle C_A,C_{BC},f=Take,d\rangle}(z\_Heroism_i = -1)$$
$$= m_{i,j=\langle C_A,C_{BC},f=Give,d\rangle}(z\_Heroism_i = +1) - m_{i,j=\langle C_A,C_{BC},f=Take,d\rangle}(z\_Heroism_i = +1)$$

| Scenario | | | Statistic | $p$-value | |
|---|---|---|---|---|---|
| $C_A$ | $C_{BC}$ | | $F(1{,}397)$ | Normal | Bonferroni |
| 0 | 0 | | 1.459 | 0.228 | 1 |
| 0 | 10 | | 0.478 | 0.490 | 1 |
| 0 | 10 | Unequal | 0.236 | 0.627 | 1 |
| 0 | 20 | | 0.078 | 0.780 | 1 |
| 10 | 0 | | 0.047 | 0.829 | 1 |
| 10 | 10 | | 0.376 | 0.540 | 1 |
| 10 | 10 | Unequal | 0.260 | 0.610 | 1 |
| 10 | 20 | | 0.081 | 0.776 | 1 |
| 20 | 0 | | 0.144 | 0.705 | 1 |
| 20 | 10 | | 0.080 | 0.778 | 1 |
| 20 | 10 | Unequal | 0.009 | 0.927 | 1 |
| 20 | 20 | | 1.486 | 0.224 | 1 |

*Notes*: Statistical significance in the last two columns is defined by * *p*<0.1; ** *p*<0.05; *** *p*<0.01

**Table 6.36.** *Hypotheses tests regarding statistically different predicted moral judgments of each scenario for high and low scores in the Deference foundation of MAC – Study 2*

$$H_0\colon m_{i,j=\langle C_A,C_{BC},f=Give,d\rangle}(z\_Deference_i = -1) - m_{i,j=\langle C_A,C_{BC},f=Take,d\rangle}(z\_Deference_i = -1)$$
$$= m_{i,j=\langle C_A,C_{BC},f=Give,d\rangle}(z\_Deference_i = +1) - m_{i,j=\langle C_A,C_{BC},f=Take,d\rangle}(z\_Deference_i = +1)$$

| Scenario | | | Statistic | $p$-value | |
|---|---|---|---|---|---|
| $C_A$ | $C_{BC}$ | | $F(1{,}397)$ | Normal | Bonferroni |
| 0 | 0 | | 0.050 | 0.822 | 1 |
| 0 | 10 | | 0.341 | 0.559 | 1 |
| 0 | 10 | Unequal | 0.818 | 0.366 | 1 |
| 0 | 20 | | 0.043 | 0.835 | 1 |
| 10 | 0 | | 0.005 | 0.944 | 1 |
| 10 | 10 | | 0.364 | 0.546 | 1 |
| 10 | 10 | Unequal | 1.326 | 0.250 | 1 |
| 10 | 20 | | 0.311 | 0.578 | 1 |
| 20 | 0 | | 0.786 | 0.376 | 1 |
| 20 | 10 | | 0.324 | 0.569 | 1 |
| 20 | 10 | Unequal | 0.201 | 0.654 | 1 |
| 20 | 20 | | 0.901 | 0.343 | 1 |

*Notes*: Statistical significance in the last two columns is defined by * *p*<0.1; ** *p*<0.05; *** *p*<0.01

**Table 6.37.** *Hypotheses tests regarding statistically different predicted moral judgments of each scenario for high and low scores in the Family foundation of MAC – Study 2*

$$H_0: m_{i,j=\langle C_A,C_{BC},f=Give,d\rangle}(z\_Family_i = -1) - m_{i,j=\langle C_A,C_{BC},f=Take,d\rangle}(z\_Family_i = -1)$$
$$= m_{i,j=\langle C_A,C_{BC},f=Give,d\rangle}(z\_Family_i = +1) - m_{i,j=\langle C_A,C_{BC},f=Take,d\rangle}(z\_Family_i = +1)$$

| Scenario | | | Statistic | $p$-value | |
|---|---|---|---|---|---|
| $C_A$ | $C_{BC}$ | | $F(1,397)$ | Normal | Bonferroni |
| 0 | 0 | | 1.512 | 0.220 | 1 |
| 0 | 10 | | 0.802 | 0.371 | 1 |
| 0 | 10 | Unequal | 0.186 | 0.667 | 1 |
| 0 | 20 | | 0.453 | 0.501 | 1 |
| 10 | 0 | | 0.182 | 0.670 | 1 |
| 10 | 10 | | 0.706 | 0.401 | 1 |
| 10 | 10 | Unequal | 0.314 | 0.576 | 1 |
| 10 | 20 | | 0.002 | 0.966 | 1 |
| 20 | 0 | | 0.948 | 0.331 | 1 |
| 20 | 10 | | 0.160 | 0.689 | 1 |
| 20 | 10 | Unequal | 0.018 | 0.892 | 1 |
| 20 | 20 | | 0.059 | 0.809 | 1 |

*Notes*: Statistical significance in the last two columns is defined by * *p*<0.1; ** *p*<0.05; *** *p*<0.01

### *A.6. Testing for significant differences of moral judgments between studies*

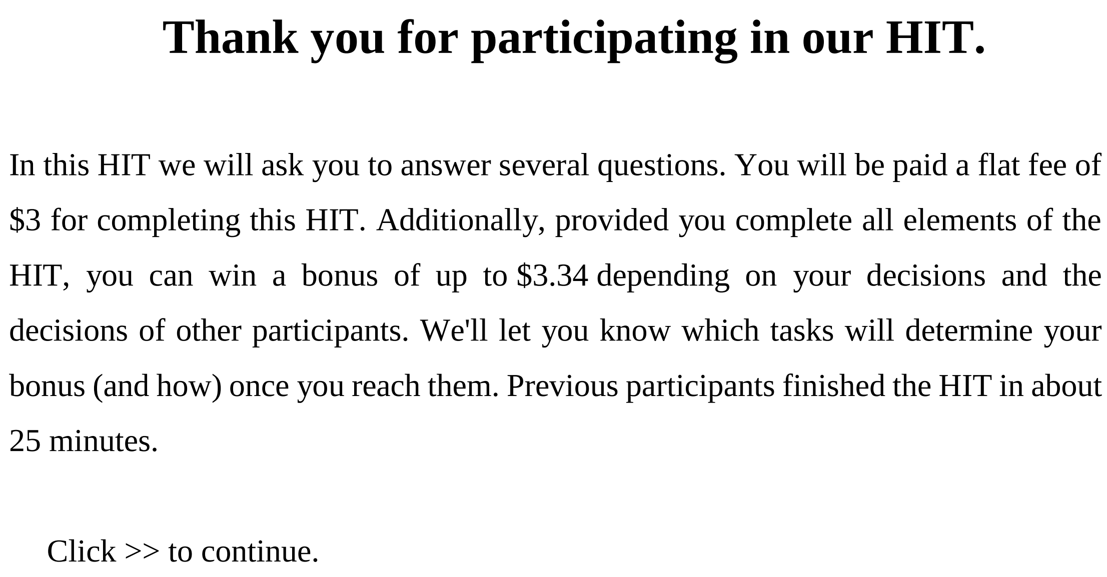
The table below provides Mann Whitney tests for the pairwise comparison of mean moral judgments of each scenario between Study 1 and Study 2. As it can be seen, no scenario has a statistically significant difference below $p < 0.05$, and none of the $p$-values survives a Bonferroni correction within a multiple comparisons' framework.

**Table 6.38.** *Testing for statistically significant differences of moral judgments of scenarios between studies*

| Scenario | | | | Mean moral judgment | | Mann Whitney |
|---|---|---|---|---|---|---|
| $C_A$ | $C_{BC}$ | $d$ | $f$ | Study 1 | Study 2 | $p$- value |
| 0 | 0 | | Give | -9.6 | -7.7 | 0.66 |
| 10 | 0 | | Give | 17.1 | 16.4 | 0.93 |
| 20 | 0 | | Give | 35.1 | 34.9 | 0.55 |
| 0 | 10 | | Give | -19.1 | -17.7 | 0.89 |
| 10 | 10 | | Give | 16.5 | 17.6 | 0.47 |
| 20 | 10 | | Give | 33.6 | 35.8 | 0.22 |
| 0 | 20 | | Give | -26.6 | -26.2 | 0.77 |
| 10 | 20 | | Give | 3.3 | 4.9 | 0.41 |
| 20 | 20 | | Give | 30.8 | 33.3 | 0.23 |
| 0 | 10 | Unequal | Give | -14 | -13.8 | 0.54 |
| 10 | 10 | Unequal | Give | 11.8 | 11.6 | 0.91 |
| 20 | 10 | Unequal | Give | 32.9 | 35.2 | 0.27 |
| 0 | 0 | | Take | -3.4 | -3.5 | 0.64 |
| 10 | 0 | | Take | 9.2 | 8.8 | 0.85 |
| 20 | 0 | | Take | 33.1 | 31.4 | 0.52 |
| 0 | 10 | | Take | -13.5 | -12.5 | 0.38 |
| 10 | 10 | | Take | 7.7 | 6.7 | 0.93 |
| 20 | 10 | | Take | 32.7 | 30.2 | 0.1* |
| 0 | 20 | | Take | -25.1 | -21.8 | 0.21 |
| 10 | 20 | | Take | -10.7 | -8.4 | 0.18 |
| 20 | 20 | | Take | 30.5 | 29.2 | 0.47 |
| 0 | 10 | Unequal | Take | -11 | -8.6 | 0.18 |
| 10 | 10 | Unequal | Take | 0.3 | 2.5 | 0.11 |
| 20 | 10 | Unequal | Take | 30.1 | 28.6 | 0.16 |

## Appendix B

### *B.1. Instructions of the Study of Chapter 3*



**Thank you for participating in our HIT.**

In this HIT we will ask you to answer several questions. You will be paid a flat fee of $3 for completing this HIT. Additionally, provided you complete all elements of the HIT, you can win a bonus of up to $3.34 depending on your decisions and the decisions of other participants. We'll let you know which tasks will determine your bonus (and how) once you reach them. Previous participants finished the HIT in about 25 minutes.

Click >> to continue.

# BEFORE YOU START!

1. Try to ensure that you will not be interrupted during the survey. You will need to complete several tasks and it is important that you take them seriously.

2. Put other devices and applications at one side so that they don't distract you.

Thank you.

# General instructions of the decision situation [*Different wording used for the Maintenance treatments in italics and between brackets*]

## Please read through the decision problem outlined below

In this decision problem, Person A will form a group with Person B. They will interact in MTurk.

To determine their bonus payment, we will first record their earnings in points and then exchange the sum of points they earned into a dollar amount for their bonus payment. Their **bonus** in Dollars will be determined as follows:

$$Earnings\ in\ Dollars = \frac{Earnings\ in\ Points}{35}$$

Person A and Person B share a **group project**. Initially, there are 0 [*60*] tokens in the project, but each person can contribute [*withdraw*] some tokens to [*from*] it. Each person has control of 30 tokens and has four options: either contribute [*withdraw*] 0, 10, 20 or 30 tokens to [*from*] the **group project.** Tokens a person does not contribute to [*withdraws from*] the project are left [*put*] by that person in his or her **private account.**

Each group member will receive an income from his or her private account and from the group project.

**<u>A group member's income from his or her private account</u>**

**A group member will receive 1 point for each token he or she leaves [*puts*] in his or her private account. No one else receives anything from tokens that he or she leaves [*puts*] in his or her private account.**

If, for example, Person A leaves [*puts*] 10 tokens in his or her private account, then person A will receive 10 points from his or her private account and Person B will receive no points from Person A's private account.

**A group member's income from the group project**

**Each group member benefits equally from tokens in the project, regardless of who put [*left*] them there.** All tokens put [*left*] in the project will be **increased by 50 percent, and the result will be split equally** among the two group members.

If, for example, Person A contributes [*withdraws*] 10 [*20*] tokens and Person B contributes [*withdraws*] 10 [*20*] tokens to [*from*] the project, then each of them will receive (10 + 10) × 1.5 / 2 = 20 × 0.75 = 15 points from the project.

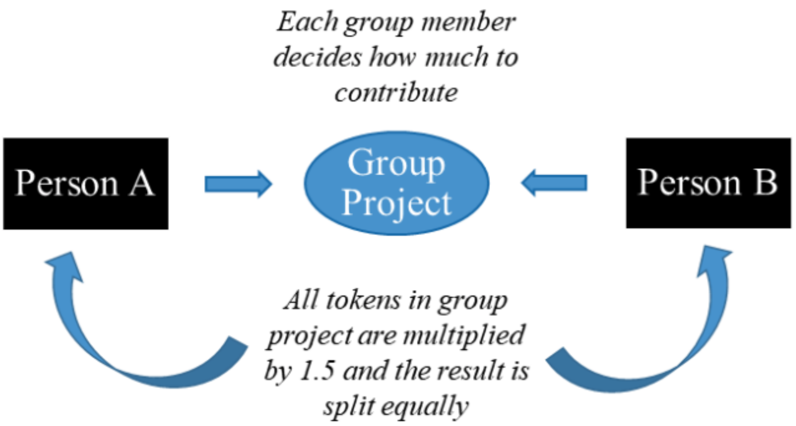
If, for example, Person A contributes [*withdraws*] 10 [*20*] tokens and Person B contributes [*withdraws*] 20 [*10*] tokens to [*from*] the project, then each of them will receive (10 + 20) × 1.5 / 2 = 30 × 0.75 = 22.5 points from the project.

**Total income**

A group member's total income is the sum of the income received from his or her private account and the income received from the group project.

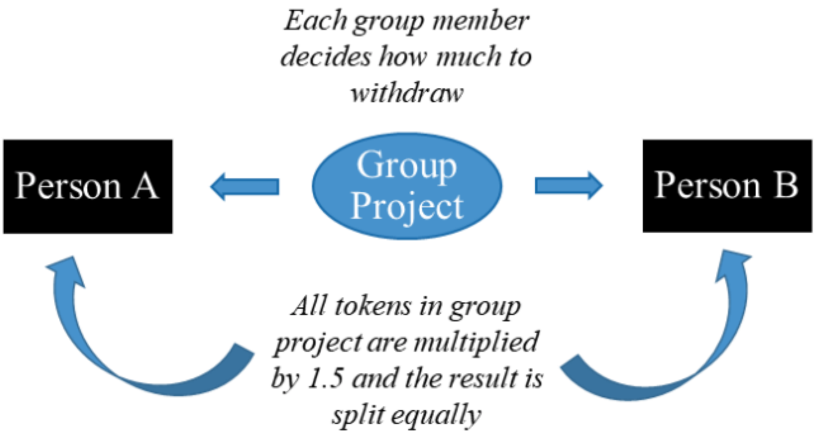
The figure below shows a summary of the interaction:

*(figure for the Provision treatments)*


Each group member decides how much to contribute
Person A
Group Project
Person B
All tokens in group project are multiplied by 1.5 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 Maintenance treatments)*


Each group member decides how much to withdraw
Person A
Group Project
Person B
All tokens in group project are multiplied by 1.5 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

**Please answer the following questions to check your understanding of the decision problem.**

## <u>Question 1.</u>

Assume that Person A contributes [*withdraws*] 0 [*30*] tokens to [*from*] the group project and Person B contributes [*withdraws*] 0 [*30*] tokens to [*from*] 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)?

Person A earnings ________

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)?

Person B earnings ________

## **Question 2.**

Assume that Person A contributes [*withdraws*] 30 [*0*] tokens to [*from*] the group project and Person B contributes [*withdraws*] 30 [*0*] tokens to [*from*] 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)?

Person A earnings ________

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)?

Person B earnings ________

## Question 3.

Assume that Person A contributes [*withdraws*] 0 [*30*] tokens to [*from*] the group project and Person B contributes [*withdraws*] 30 [*0*] tokens to [*from*] 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)?

Person A earnings ________

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)?

Person B earnings ________

## Question 4.

Assume that Person A contributes [*withdraws*] 30 [*0*] tokens to [*from*] the group project and Person B contributes [*withdraws*] 0 [*30*] tokens to [*from*] 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)?

Person A earnings ________

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)?

Person B earnings ________

## **Question 5.**

Assume that Person A contributes [*withdraws*] 20 [*10*] tokens to [*from*] the group project and Person B contributes [*withdraws*] 10 [*20*] tokens to [*from*] 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)?

Person A earnings ________

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)?

Person B earnings ________

# General introduction to the P- and C-experiments

We will now present you **two decision situations**. Your overall bonus from the HIT will depend on the decisions you and the other group member (another MTurker) take in various tasks that follow. For each task, we will explain how your and the other group member 's decisions will affect your bonus.

## How your final bonus from this HIT will be determined

Your total bonus is computed by adding the bonus you earn in each of the tasks presented in the next two decision situations. Remember that the points will be exchanged to dollars as described in the decision problem presented at the beginning of the HIT.

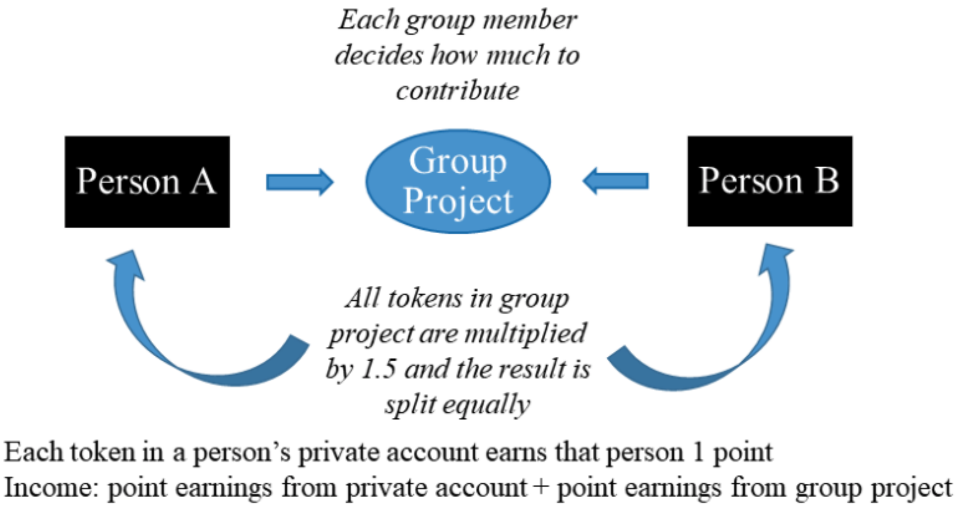
Press continue when you are ready.

# P-experiment

In this decision situation, **you will form a group with another person. You will interact in MTurk**.

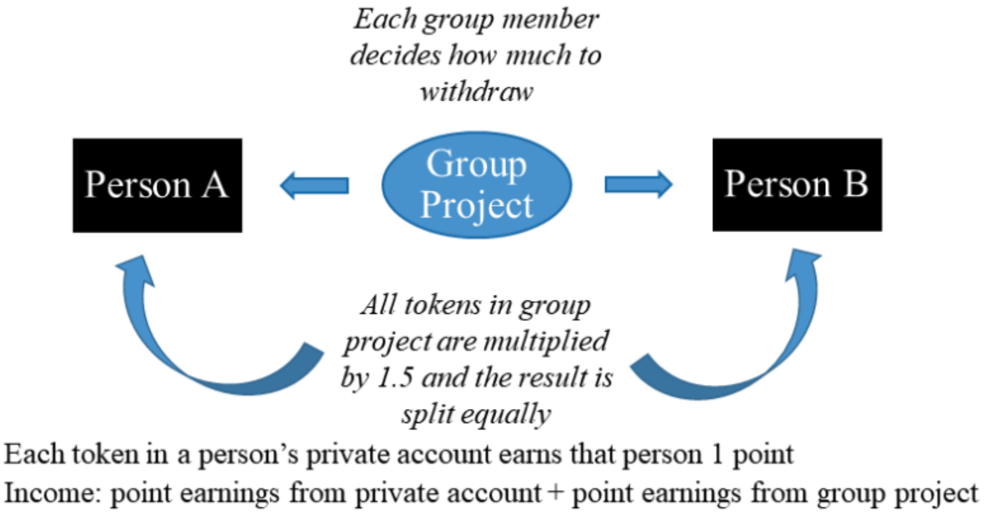
Your tasks here are based on the decision problem described at the beginning of the HIT, which is summarised in the following figure:

*(figure for the Provision treatments)*


*Each group member decides how much to contribute*
Person A
Group Project
Person B
*All tokens in group project are multiplied by 1.5 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 Maintenance treatments)*


*Each group member decides how much to withdraw*
Person A
Group Project
Person B
*All tokens in group project are multiplied by 1.5 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

All group members have **two** tasks, which we will refer to below as the "**unconditional contribution [*withdrawal*]**" and the "**contribution [*withdrawal*] table**".

In the **unconditional contribution** [*withdrawal*] task you simply decide the amount of tokens (either 0, 10, 20 or 30) you want to contribute to [withdraw from] the group project.

In the **contribution [*withdrawal*] table** task you indicate the amount of tokens **you want to contribute to [*withdraw from*]** the group project **for each possible contribution [*withdrawal*] of the other group member**. Here, you can condition your contribution [*withdrawal*] on that of the other group member.

This is a one-off situation that is finished once you have made both decisions.

**<u>How your bonus from this decision situation will be determined</u>**

We will randomly select one group member for whom the unconditional contribution [*withdrawal*] will be relevant for their earnings once you and the other group member have made your decisions. The contribution [*withdrawal*] table will determine the earnings of the non-selected group member.

**Example:**

- Imagine that the unconditional contributions [*withdrawals*] of group members A and B are 20 [*10*] and 10 [*20*], respectively.
- Assume that group member A has been randomly selected so that his or her unconditional contribution [*withdrawal*] (20 [*10*] in this example) is

relevant for his or her earnings. Hence, group member B's contribution [*withdrawal*] table will be used to calculate his or her earnings.

- To determine the contribution [*withdrawal*] of group member B we will take the contribution [*withdrawal*] this group member indicates in their contribution [*withdrawal*] table if group member A contributes 20 [*withdraws 10*].
- Imagine that group member B contributes 30 [*withdraws 0*] if group member A contributes 20 [*withdraws 10*]. Then the total sum of contributions to [*tokens left in*] the group project are 20 + 30 = 50 tokens.
- Hence, group member A earns 10 + 50 × 1.5/2 = 47.5 points and group member B earns 0 + 50 × 1.5/2 = 37.5 points.

Press continue when you are ready.

## The unconditional contribution [*withdrawal*]

How many tokens out of 30 do you contribute to [*withdraw from*] the group project, i.e. 0, 10, 20 or 30?

## **The contribution [*withdrawal*] table**

Now we ask you to think about your contribution [*withdrawal*] depending on how much the other group member contributes [*withdraws*]. Please indicate for each possible contribution [*withdrawal*] of the other group member how much you contribute [*withdraw*], i.e. 0, 10, 20 or 30.

| | I contribute [*withdraw*] |
|---|---|
| If other contributes [*withdraws*] 0 [*30*] | |
| If other contributes [*withdraws*] 10 [*20*] | |
| If other contributes [*withdraws*] 20 [*10*] | |
| If other contributes [*withdraws*] 30 [*0*] | |

# C-experiment

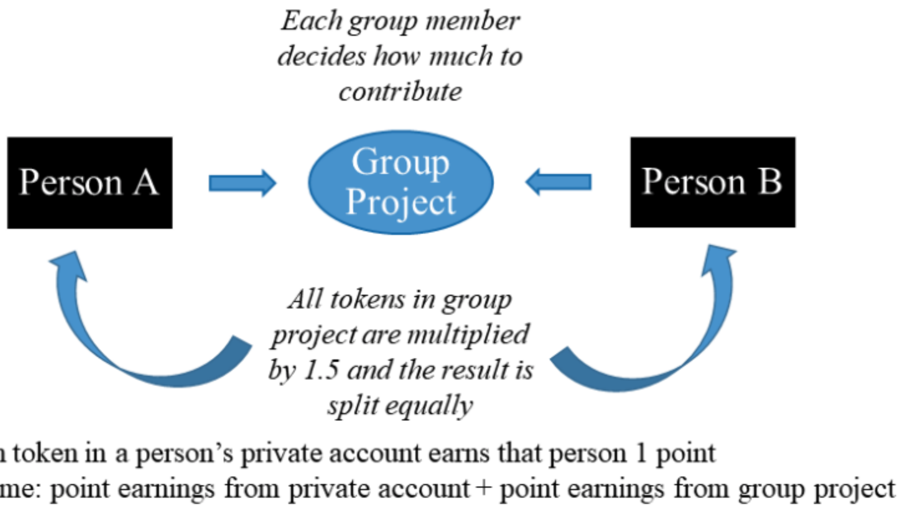
Please now consider another decision situation, consisting of two decision tasks.

In this decision situation, **you will form a group with another person. You will interact in MTurk**. **The MTurker you will be paired with in this decision**

**situation is a different one from the MTurker you were paired with in the previous decision situation.**

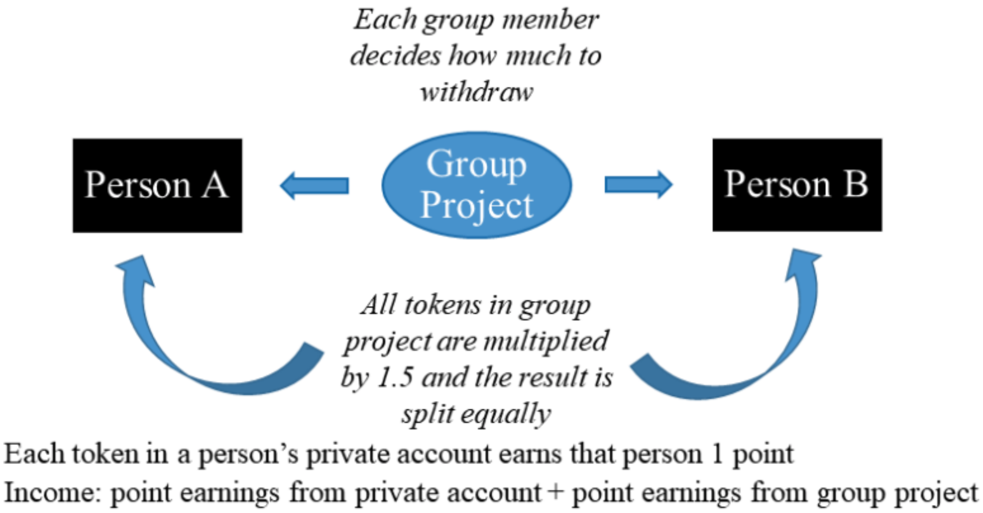
Your tasks here are based on the decision problem described at the beginning of the HIT, which is summarised in the following figure:

*(figure for the Provision treatments)*


Each group member decides how much to contribute
Person A
Group Project
Person B
All tokens in group project are multiplied by 1.5 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 Maintenance treatments)*


Each group member decides how much to withdraw
Person A
Group Project
Person B
All tokens in group project are multiplied by 1.5 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

All group members have **two** tasks, which we will refer to below as the "**contribution [*withdrawal*] task**" and the "**prediction task**".

In the **contribution [*withdrawal*] task** you have to decide the amount of tokens (either 0, 10, 20 or 30) you want to contribute to [withdraw from] the project.

## How the bonus from the contribution [*withdrawal*] task will be determined

**The bonus you earn from this task** is determined as explained in the decision problem presented at the beginning of the HIT. **Your decision and the decision of the other group member will determine the tokens left [*put*] in your corresponding private accounts and the total amount of tokens put [*left*] in the group project.**

In the **prediction task** you have to predict the contribution to [*withdrawal from*] the group project of the other group member.

## How the bonus from the prediction task will be determined

Your bonus will depend on the accuracy of your prediction:

- If your prediction is exactly right (that is, if your prediction is **exactly** the same as the actual contribution [*withdrawal*] of the other group member), you will get **12 points** in addition to the points you earn in the other decision task.
- Otherwise, you will not get any additional points.

## How your bonus from this decision situation will be determined

Your **bonus from this decision situation** is computed by adding the bonus you earn in the **contribution [*withdrawal*] task** and the **prediction task**.

Press continue when you are ready.

How many tokens out of 30 do you contribute to [*withdraw from*] the group project, i.e. 0, 10, 20 or 30?

What is your prediction of the contribution to [*withdrawal from*] the group project of the other group member, i.e. 0, 10, 20 or 30?

## General Introduction to the M-experiment and the two moral questionnaires (MFQ and MAC-Q)

The goal of the following tasks is to investigate people's views of various social, moral and political issues. You will also answer some questions about yourself at some point. These tasks will be presented in the next 12 screens.

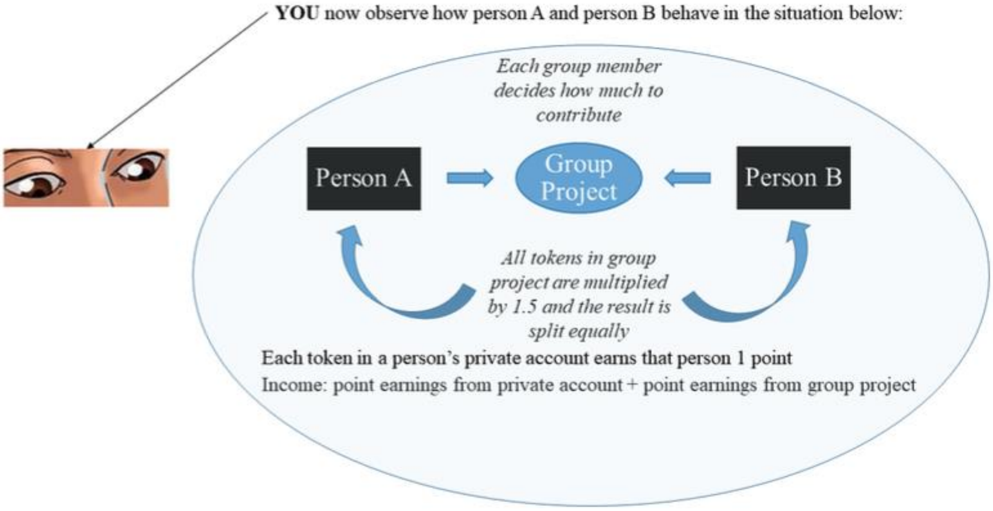
There are no correct or incorrect answers -- just respond in a way that feels appropriate to you. Do not think too long about any question, your first answer is probably the best. And if you are unsure what a statement means, just give an answer that fits with your understanding of it.

Press continue when you are ready.

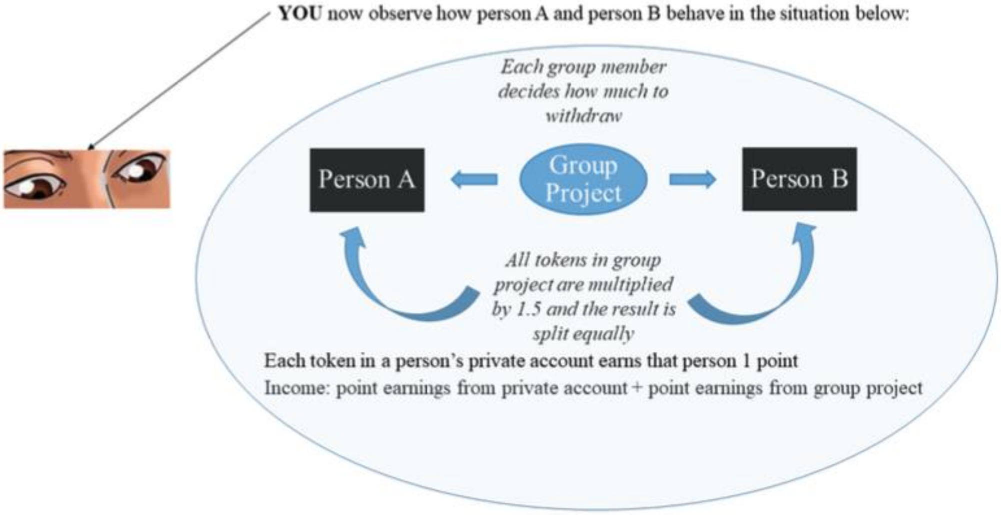
## M-experiment

**You are now an outside OBSERVER of the decision problem** presented at the beginning of the HIT.

*(figure for the Provision treatments)*


YOU now 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 1.5 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

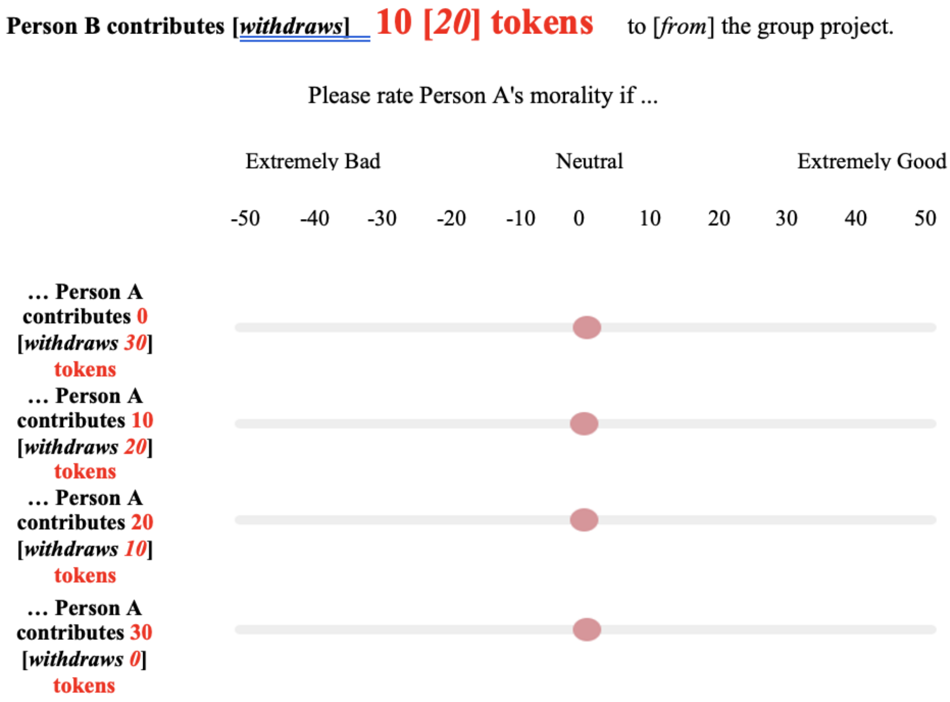
*(figure for the Maintenance treatments)*


YOU now observe how person A and person B behave in the situation below:
Each group member decides how much to withdraw
Person A
Group Project
Person B
All tokens in group project are multiplied by 1.5 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.

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.

*[We displayed the scenarios in 4 different screens, four scenarios per each screen. The four scenarios per screen fixed what Person B C did in a line summarising their actions. Each scenario per screen varied what Person A did. The scenarios displayed below have been worked so that they resemble how the screens looked to the participant]*



**Person B contributes [*withdraws*] 10 [*20*] tokens** to [*from*] the group project.

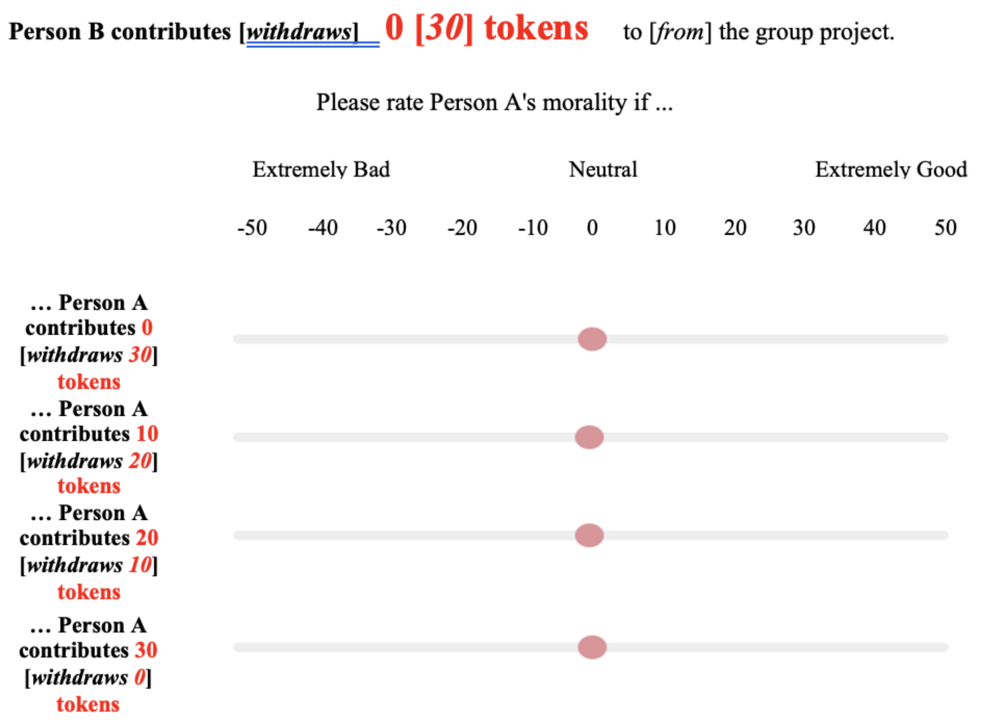
Please rate Person A's morality if ...

**... Person A contributes 0 [*withdraws 30*] tokens**

**... Person A contributes 10 [*withdraws 20*] tokens**

**... Person A contributes 20 [*withdraws 10*] tokens**

**... Person A contributes 30 [*withdraws 0*] tokens**



**Person B contributes [*withdraws*] 0 [*30*] tokens** to [*from*] the group project.

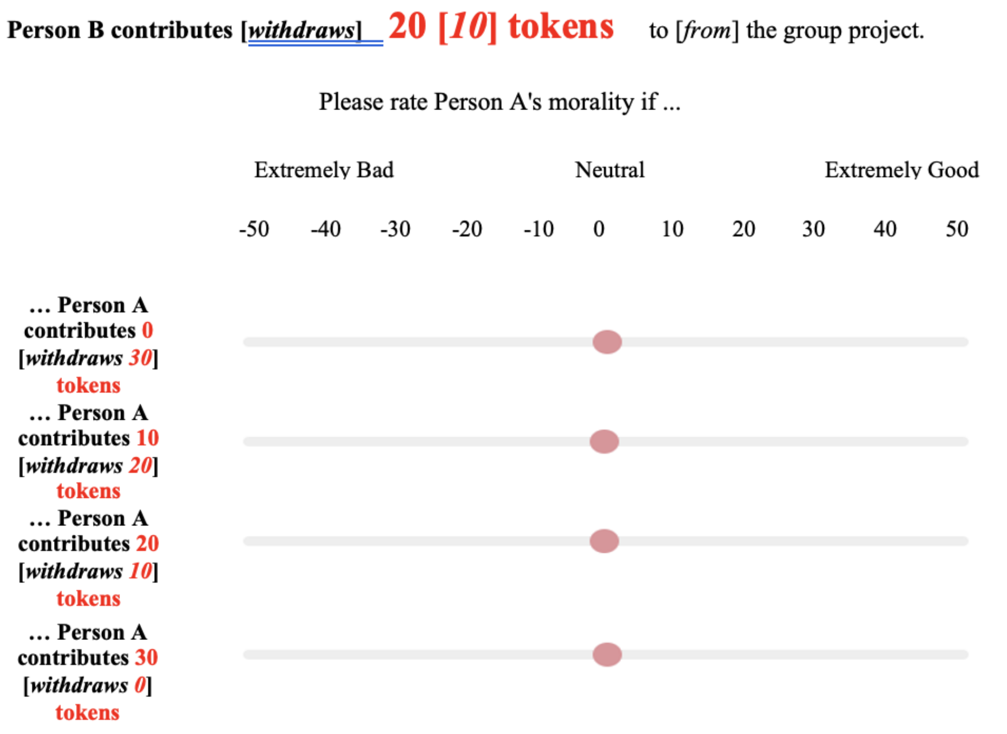
Please rate Person A's morality if ...

**… Person A contributes 0 [*withdraws 30*] tokens**

**… Person A contributes 10 [*withdraws 20*] tokens**

**… Person A contributes 20 [*withdraws 10*] tokens**

**… Person A contributes 30 [*withdraws 0*] tokens**



**Person B contributes [*withdraws*] 20 [*10*] tokens** to [*from*] the group project.

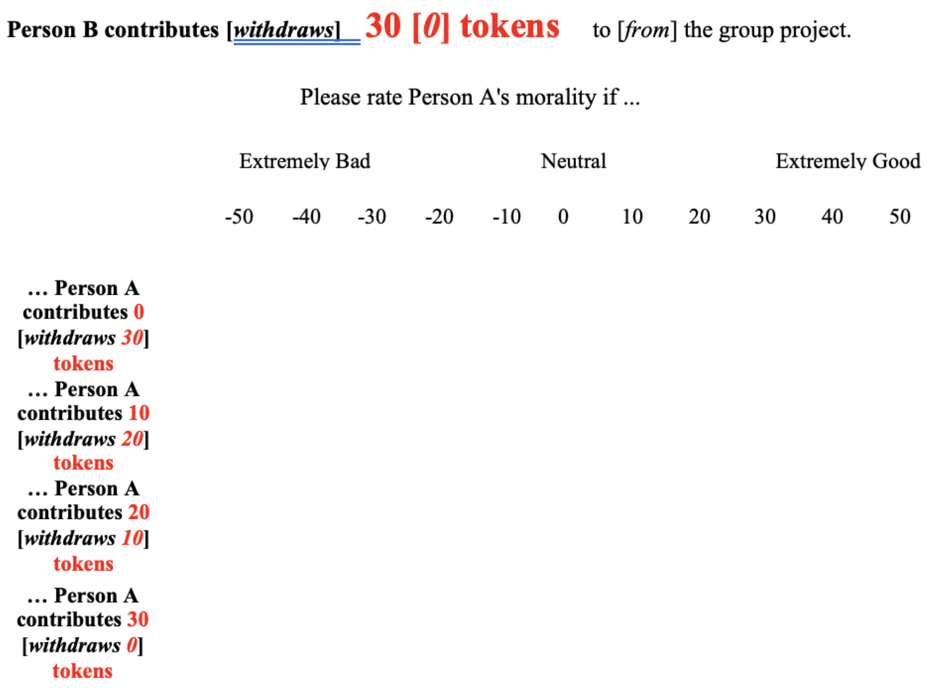
Please rate Person A's morality if ...

**… Person A contributes 0 [*withdraws 30*] tokens**

**… Person A contributes 10 [*withdraws 20*] tokens**

**… Person A contributes 20 [*withdraws 10*] tokens**

**… Person A contributes 30 [*withdraws 0*] tokens**



**Person B contributes [*withdraws*] 30 [*0*] tokens** to [*from*] 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 [*withdraws 30*] tokens** | | | | | | | | | | | |
| **… Person A contributes 10 [*withdraws 20*] tokens** | | | | | | | | | | | |
| **… Person A contributes 20 [*withdraws 10*] tokens** | | | | | | | | | | | |
| **… Person A contributes 30 [*withdraws 0*] tokens** | | | | | | | | | | | |

# Sociodemographic questionnaire

**Thank you for finishing the previous part. Now, please read through the questions below and answer them as accurately as possible.**

*[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.** How many hours in total do you work per week?
*[Options to the respondent: 0 – 20, 20 – 40, 40 – 60, 60 – 80, More than 80]*

**Q2.** Your Gender:
*[Options to the respondent: Male, Female, Prefer not to say]*

**Q3.** Your Age:
*[Options to the respondent: from 15 to 100 in steps of 1]*

**Q4.** What is your nationality?
*[Options to the respondent: The default list of countries]*

**Q5.** Would you describe yourself as a liberal, conservative or something else?

*[Options to the respondent: Moderate/Middle of the Road, Liberal, Very Liberal, Conservative, Very Conservative, Libertarian, Other, Prefer not to say]*

**Q6.** How religious are you?

*[Options to the respondent: Not at all, Somewhat religious, Very religious, Prefer not to say]*

**Q7.** 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]*

**Q8.** What is your highest qualification attained?

*[Options to the respondent: Less than high school, High school, Vocational Training, Attended University but didn't finish, Undergraduate Degree, Postgraduate Degree, Prefer not to say]*

**Q9.** Please choose the category that describes the total amount of income you earned this year.

*[Options to the respondent: $5,000 or less, $5,001 – $25,000, $25,001 – $50,000, $50,001 – $75,000, $75,001 – $100,000, More than $100,000, Prefer not to say]*

**Q10.** 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]*

# Moral Foundations Questionnaire (MFQ)

**Thank you for finishing the previous part. Now, please answer the following questionnaire.**

**Part 1.** When you decide whether something is right or wrong, to what extent are the following considerations relevant to your thinking? Please answer on a scale from

*Not At All Relevant* (this consideration has nothing to do with my judgments of right and wrong) to *Extremely Relevant* (this is one of the most important factors when I judge right and wrong)

*[Options for the respondent: Not At All Relevant, Not Very Relevant, Slightly Relevant, Somewhat Relevant, Very Relevant, Extremely Relevant]*

*[Each sentence was displayed with Font Times New Roman, size 18, bold and centred. The options for the respondent appeared below each of the statements]*

Whether or not someone conformed to the traditions of society.
Whether or not an action caused chaos or disorder.
Whether or not some people were treated differently than others.
Whether or not someone was denied his or her rights.
Whether or not someone acted in a way that God would approve of.
Whether or not someone showed a lack of respect for authority.
Whether or not someone was good at math.
Whether or not someone's action showed love for his or her country.
Whether or not someone cared for someone weak or vulnerable.
Whether or not someone suffered emotionally.
Whether or not someone showed a lack of loyalty.
Whether or not someone did something disgusting.
Whether or not someone did something to betray his or her group.
Whether or not someone acted unfairly.
Whether or not someone violated standards of purity and decency.
Whether or not someone was cruel.

**Part 2.** Please read the following sentences and indicate your level of agreement or disagreement

*[Options for the respondent: Strongly Disagree, Moderately Disagree, Slightly Disagree, Slightly Agree, Moderately Agree, Strongly Agree]*

*[Each sentence was displayed with Font Times New Roman, size 18, bold and centred. The options for the respondent appeared below each of the statements]*

It is more important to be a team player than to express oneself.

Men and Women have different roles to play in society.

Chastity is an important and valuable virtue.

I am proud of my country's history.

Justice is the most important requirement for a society.

I would call some acts wrong on the grounds that they are unnatural.

Compassion for those who are suffering is the most crucial virtue.

It is better to do good than to do bad.

One of the worst things a person could do is hurt a defenseless animal.

People should be loyal to their family members, even when they have done something wrong.

When the government makes laws, the number one principle should be ensuring that everyone is treated fairly.

I think it's morally wrong that rich children inherit a lot of money while poor children inherit nothing.

Respect for authority is something all children need to learn.

People should not do things that are disgusting, even if no one is harmed.

If I were a soldier and disagreed with my commanding officer's orders, I would obey anyway because that is my duty.

It can never be right to kill a human being.

## Morality As Cooperation Questionnaire (MAC-Q)

**Thank you for finishing the previous part. Now, please answer the following questionnaire.**

*[MAC-Q is displayed differently (and on a different scale) than the MFQ. We preserve how MAC-Q was displayed in Curry et al (2019). To do that, we displayed each of the statements in a slider frame (as the moral evaluations of scenarios)]*

**When you decide whether something is right or wrong, to what extent are the following considerations relevant to your thinking?**

Click on the lines below, and/or move the sliders

*[Options for the respondent: Continuous scale via slider. Guide above the sliders showed the following guide: Not At All Relevant, Not Very Relevant, Slightly Relevant, Somewhat Relevant, Very Relevant, Extremely Relevant]*

Whether or not someone acted to protect their family

Whether or not someone helped a member of their family

Whether or not someone’s action showed love for their family

Whether or  not someone acted in a way that helped their community

Whether or not someone helped a member of their community

Whether or not someone worked to unite a community

Whether or not someone did what they had agreed to do

Whether or not someone kept their promise

Whether or not someone proved that they could be trusted

Whether or not someone acted heroically

Whether or not someone showed courage in the face of adversity

Whether or not someone was brave

Whether or not someone deferred to those in authority

Whether or not someone disobeyed orders

Whether or not someone showed respect for authority

Whether or not someone kept the best part for themselves

Whether or not someone showed favouritism

Whether or not someone took more than others

Whether or not someone vandalised another person's property

Whether or not someone kept something that didn't belong to them

Whether or not someone's property was damaged

## To what extent do you agree with the following statements?

Click on the lines below, and/or move the sliders

*[Continuous scale via slider. Guide above the sliders showed the following guide: Strongly Disagree, Disagree, Neither Agree nor Disagree, Agree, Strongly Agree]*

People should be willing to do anything to help a member of their family

You should always be loyal to your family

You should always put the interests of your family first

People have an obligation to help members of their community

It's important for individuals to play an active role in their communities

You should try to be a useful member of society

You have an obligation to help those who have helped you

You should always make amends for the things you have done wrong

You should always return a favour if you can

Courage in the face of adversity is the most admirable trait

Society should do more to honour its heroes

To be willing to lay down your life for your country is the height of bravery

People should always defer to their superiors

Society would be better if people were more obedient to authority

You should respect people who are older than you

Everyone should be treated the same

Everyone’s rights are equally important

The current levels of inequality in society are unfair

It’s acceptable to steal food if you are starving

It’s ok to keep valuable items that you find, rather than try to locate the rightful owner

Sometimes you are entitled to take things you need from other people

# Final Comment

**That's the end of the study. Thank you very much for taking part!**

If you want to leave any comments about how did you make your decisions in the bonus-related tasks, please leave them in the box below.

# Appendix C

## *C.1. Instructions of the Study of Chapter 4*

**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**

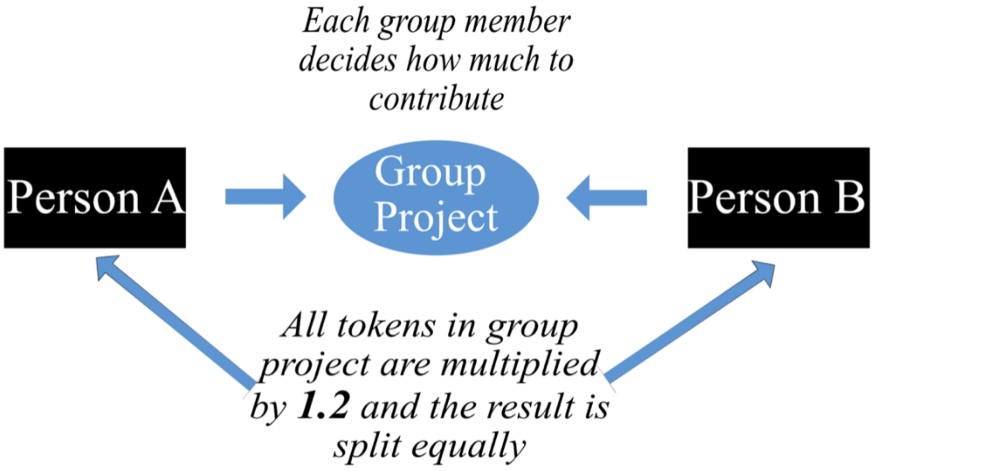
**Each person benefits equally from tokens in the group project, regardless of who put them there.** All tokens put in the project will be **multiplied by 1.2 [2.4], and the result will be split equally** among the two persons interacting.

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**

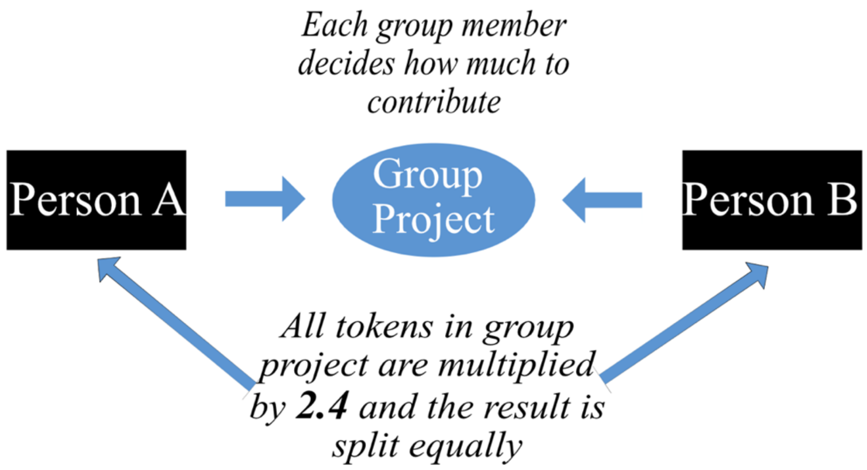
Each person receives the income from their own private account plus their share of income from the group project.

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)?

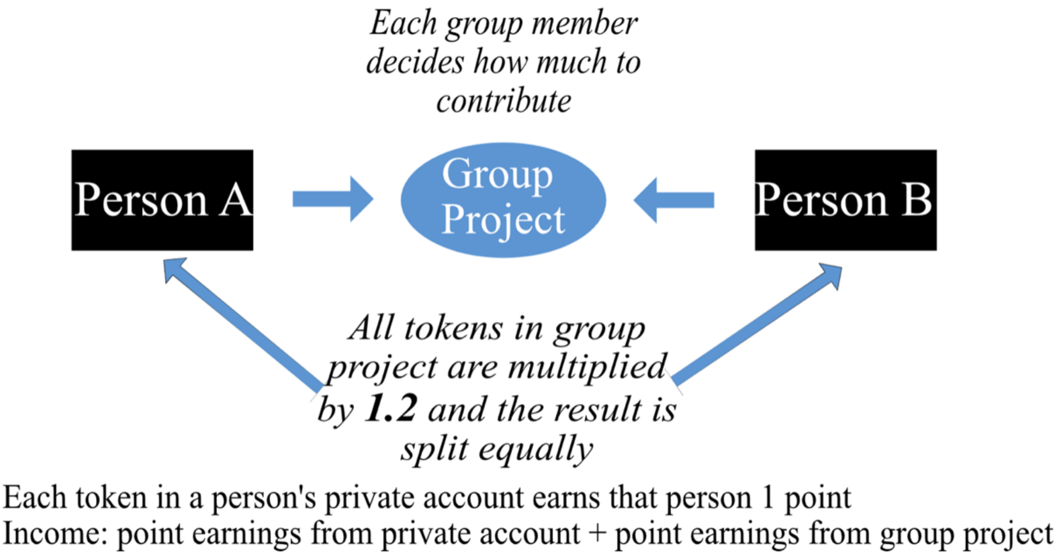
______________________________________________________________

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)?

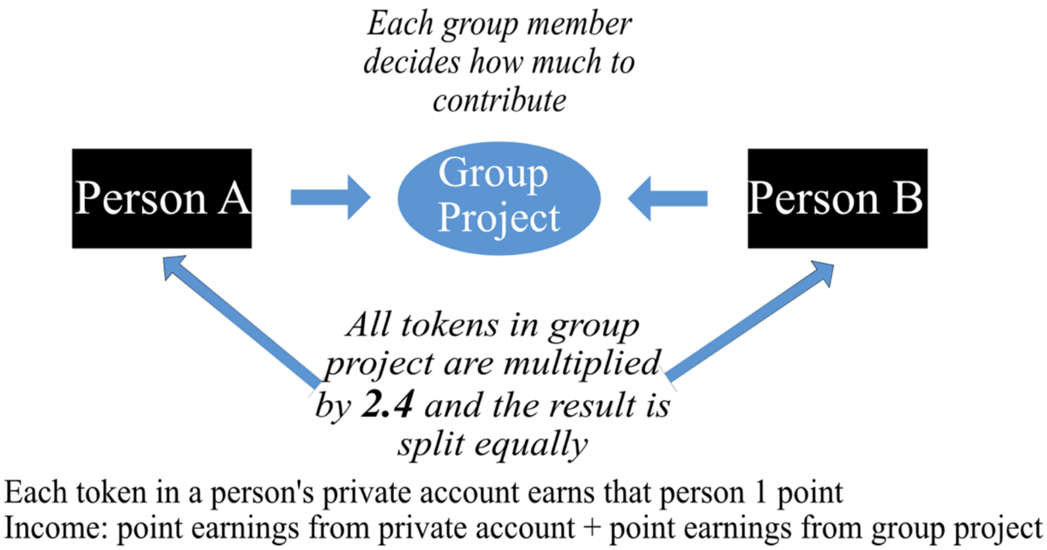
______________________________________________________________

**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 | |

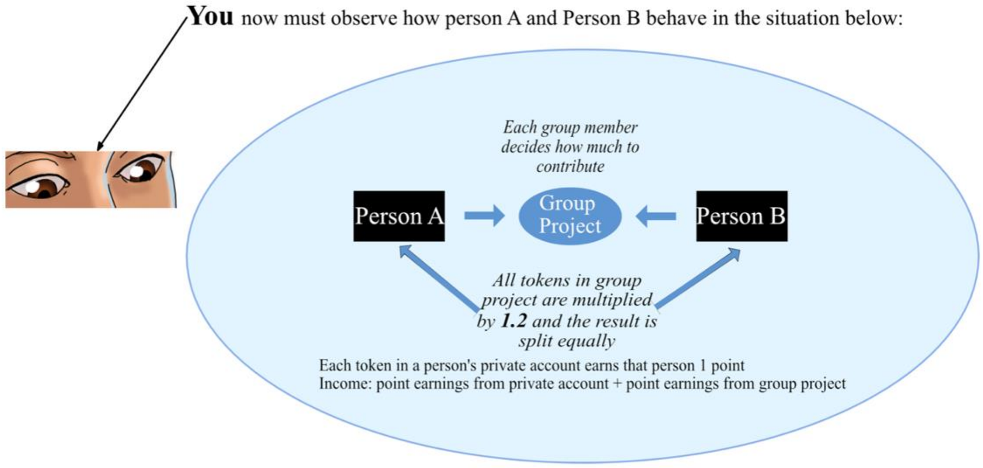
**Instructions for the M-experiment**

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.

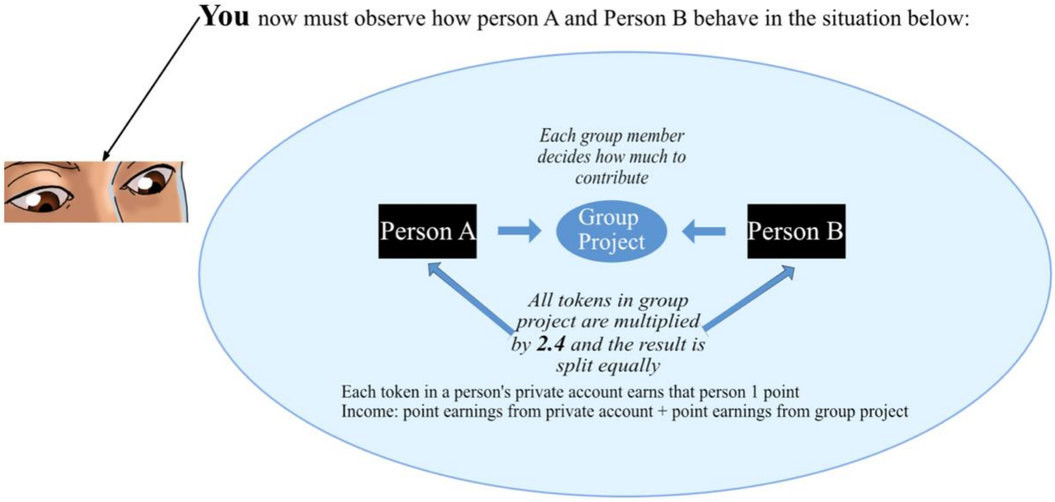
There are no correct or incorrect answers - just respond with what **you really think**

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
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

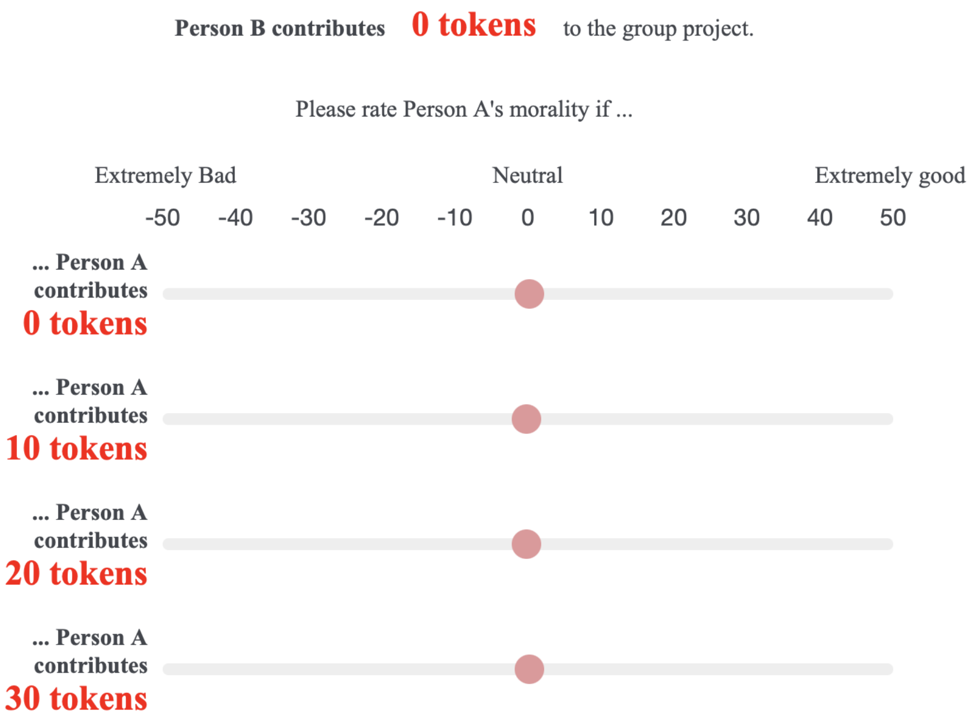
*(figure for the social dilemma game)*


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 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.

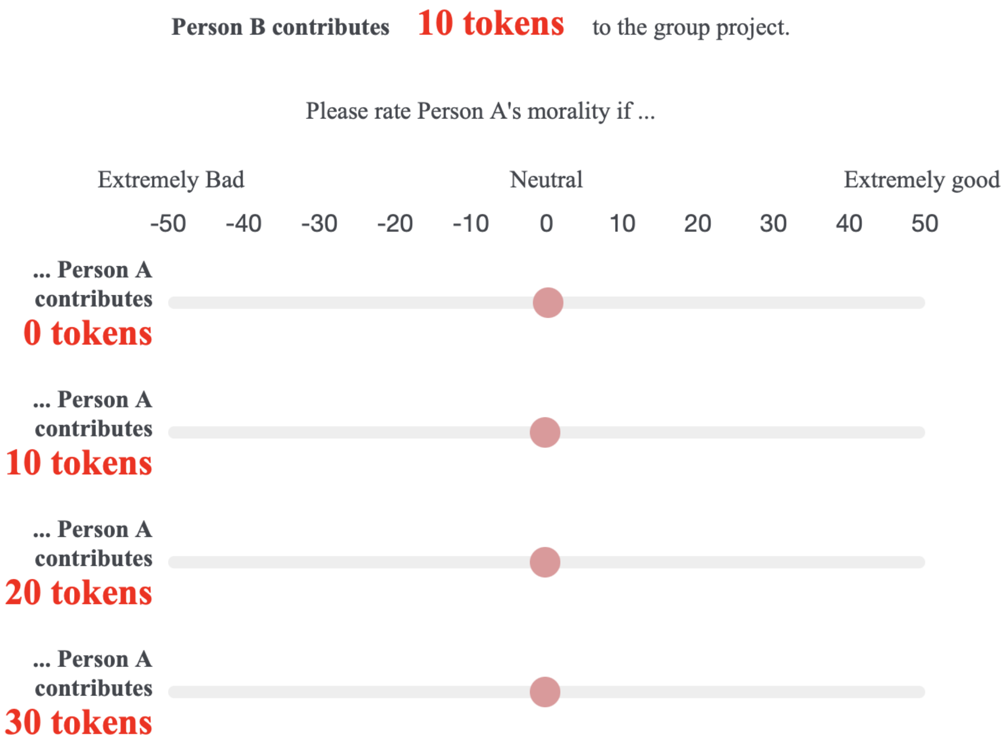
Please rate Person A's morality if ...

**... 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.

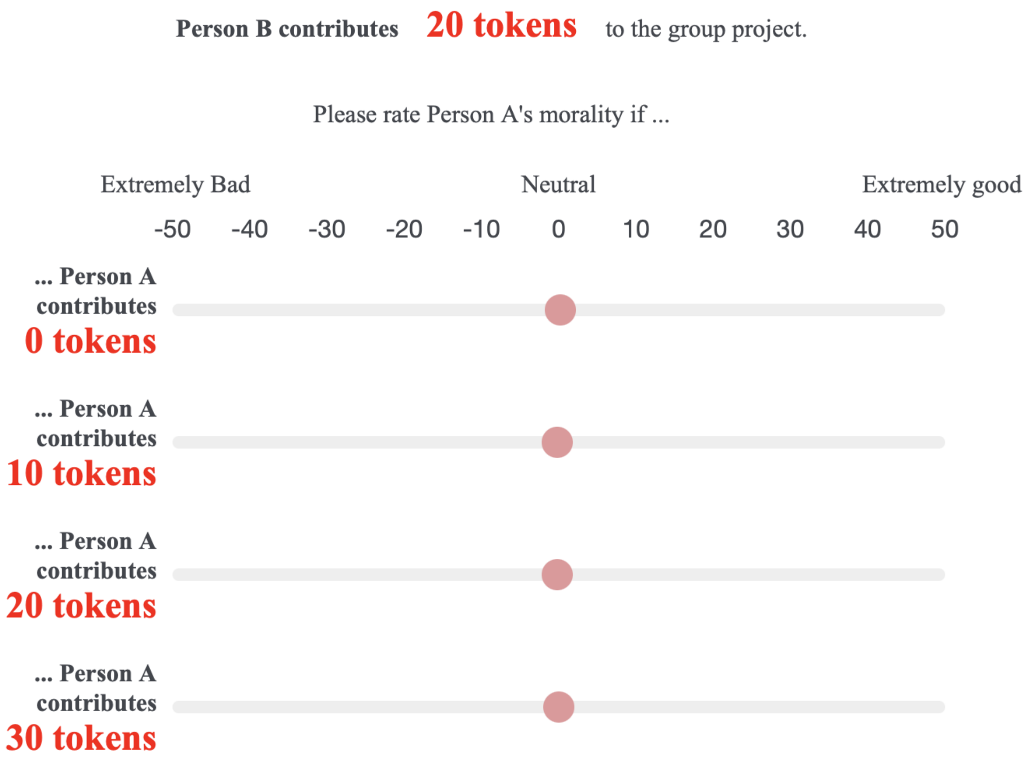
Please rate Person A's morality if ...

**... 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 ...

Extremely Bad Neutral Extremely good

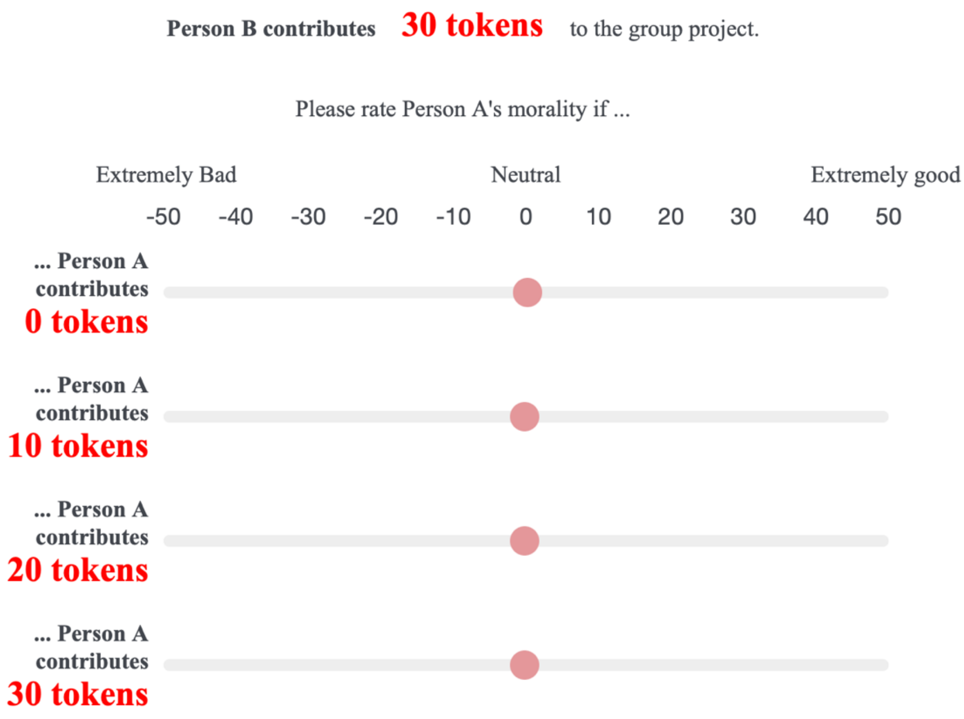
-50 -40 -30 -20 -10 0 10 20 30 40 50

**... Person A contributes 0 tokens**

**... Person A contributes 10 tokens**

**... Person A contributes 20 tokens**

**... Person A contributes 30 tokens**



**Person B contributes** **30 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**

## 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 <u>same in all decision situations</u>. *Distribution 1* is <u>different in all decision situations</u>.**

**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

________________________________________

## *C.2. Theoretical appendix of Chapter 4*

### C.2.1. *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, $\overline{m} \in (1, \infty)$).

### *C2.2. The proofs*

#### *C.2.2.1. Predictions of theories regarding contribution preferences*

### C.2.2.1.1. 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.*

### C.2.2.1.2. 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.*

C.2.2.1.3. Inequality Aversion Preferences

### *C.2.2.1.3.1. 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.*

*<u>First part of the proof: proving (i)</u>*

*Step 1: Recall necessary functions.*

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

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

*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, $U_i^{FS}(\pi_i, \pi_j)$ 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 $\pi_i(\bar{g}_i, g_j) = 30 - \bar{g}_i + m \times (\bar{g}_i + g_j)$ in the first term of the RHS and using, again, the results from lemma 0 (b), we can rewrite the utility function as follows:

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

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

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

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

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

*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 $\alpha_i$ 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 $\beta_i$ 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 $\left(\forall \langle g_i g_j \rangle \in A_i \times A_j\right), \frac{\partial U_i^{FS}\left(\pi_i(g_i,g_j),\pi_j(g_i,g_j)\right)}{\partial g_i} = 0 \; iff g_i \in \left[0, g_j\right]$; 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 $g_i = g_j$.

*Second part of the proof: proving (ii)*

*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 $\beta_i$* 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* $\alpha_i$ *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 $\left(\forall \langle g_i g_j \rangle \in A_i \times A_j\right), \frac{\partial U_i^{FS}\left(\pi_i(g_i, g_j), \pi_j(g_i, g_j)\right)}{\partial g_i} = 0 \; iff \; g_i \in \left[g_j, 30\right]$, 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.*

#### *C.2.2.1.3.2. 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.*

#### C.2.2.1.4. Reciprocity preferences

#### C.2.2.1.4.1. *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 ($\kappa_{ij}$) 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:

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

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

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

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 ($\lambda_{iji}$) 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:

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

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

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

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:

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

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

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

Which simplify to:

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

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

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

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:

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

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:

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

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

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

Which can be further simplified to:

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

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:

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

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

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

*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:

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

*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:

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

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 ($\kappa_{ij}$) and the perceived kindness function ($\lambda_{iji}$) 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,

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

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:

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

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

Which can be simplified to:

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

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

*Step 11: find the kindness function ($\kappa_{ij}$) 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:

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

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

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

Simplifying, we get:

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

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 ($\lambda_{iji}$) 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:

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

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

Which simplify to:

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

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

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:

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

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

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

Which can be further simplified to:

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

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

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

Which can be further simplified to:

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

*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:

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

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:

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

*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:

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

*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

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

*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:

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

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

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

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

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

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

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

*QED.*

#### *C.2.2.1.4.3. 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:

$$c_i^* = g_i = 0 \text{ against } g_j = 20 \text{ iff } Y_{i,j} < \frac{1-0.6}{0.36\times(5)}$$
$$c_i^* = g_i = 30 \text{ against } g_j = 20 \text{ iff } Y_{i,j} > \frac{1-0.6}{0.36\times(5)}$$
$$c_i^* = g_i = 0 \text{ against } g_j = 30 \text{ iff } Y_{i,j} < \frac{1-0.6}{0.36\times(15)}$$
$$c_i^* = g_i = 30 \text{ against } g_j = 30 \text{ iff } Y_{i,j} > \frac{1-0.6}{0.36\times(15)}$$

And, simplifying, we get:

$$c_i^* = g_i = 0 \text{ against } g_j = 20 \text{ iff } Y_{i,j} < \frac{0.4}{1.8}$$
$$c_i^* = g_i = 30 \text{ against } g_j = 20 \text{ iff } Y_{i,j} > \frac{0.4}{1.8}$$
$$c_i^* = g_i = 0 \text{ against } g_j = 30 \text{ iff } Y_{i,j} < \frac{0{,}4}{5.4}$$
$$c_i^* = g_i = 30 \text{ against } g_j = 30 \text{ iff } Y_{i,j} > \frac{0.4}{5.4}$$

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:

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

Substituting $g_j$ explicitly in the inequalities, we get:

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

$c_i^* = g_i = 0$ 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 $Y_{i,j}$:

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

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 $0.8\times(>1) = (>0.8)$. 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.*

C.2.2.1.5. Spiteful preferences

### *C.2.2.1.5.1. 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:

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

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:

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

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 $g_i \ge 0$. Hence, against $g_j = 0$ the best response is to fully contribute regardless of the value of $\beta_i$, 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

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

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 $\beta_i$ 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 $\beta_i$ 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.*

#### *C.2.2.1.5.2. Other results involving spiteful preferences*

Below we provide a corollary that presents the specific threshold values of $\beta_i$ 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.*

C.2.2.1.6. Social Efficiency preferences

### *C.2.2.1.6.1. 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.*

### *C.2.2.1.6.2. 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*

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

*(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:

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

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

*QED.*

### C.2.2.1.7. Maximin preferences

### *C.2.2.1.7.1. 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.*

### *C.2.2.1.7.2. 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.*

#### *C.2.2.2. Proofs regarding estimated parameters through the use of parameter-elicitation games*

### C.2.2.2.1. 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

- $\varepsilon$ 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 $\pi_j(x, 14 - x) = 14 - x$.

*C.2.2.2.1.1. Disadvantageous Inequality parameter*

C.2.2.2.1.1.1. 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 $\alpha_i$ 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 $\varepsilon$. 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 $\alpha_i$, 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.*

#### C.2.2.2.1.1.2. 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 $\alpha_i$ 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.*

### C.2.2.2.2. 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.

- $\varepsilon$ 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

*C.2.2.2.2.1. Advantageous Inequality and Spiteful parameters*

C.2.2.2.2.1.1. 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 $\beta_i$ 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 $\beta_i$ 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, $\beta_i$ can be expressed in terms of the two thresholds together:

$$\frac{20-(x+\varepsilon+\varepsilon)}{20} < \beta_i < \frac{20-(x+\varepsilon)}{20}$$

*QED.*

#### C.2.2.2.2.1.2. 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 $\beta_i$ 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≥ 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 $\beta_i$ 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:

$$MRQ := \left\{x \in MRQ \middle| \left((x \in X) \wedge \left(x < \frac{-(\geq 4)}{20}\right)\right)\right\}$$

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 $\beta_i$ 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.*

*C.2.2.2.2.2. Social Efficiency parameter*

C.2.2.2.2.2.1. Proof of proposition 9.

**Proposition 9.** *Let's suppose that subject $i$ maximizes the utility function $U_i^{SE}\left(\pi_i, \pi_j\right)$. 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.*

#### C.2.2.2.2.2.2. 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.*

*C.2.2.2.2.3. Maximim parameter*

C.2.2.2.2.3.1. 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 $\beta_i$.*

*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 $\beta_i$ 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 $\beta_i$ will be the same, which proves (b).

*QED.*

C.2.2.2.2.3.2. 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 $\beta_i$ 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.*

### C.2.2.2.3. 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 $x_2 < x_1$.

### *C.2.2.2.3.1. Reciprocity parameter - Proof of proposition* 11.

**Proposition 11.** *Let's suppose that subject $i$ maximizes the utility function $U_i^{DK}(\pi_i, \pi_j)$. 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:

$$\begin{aligned}\kappa_i(a_{ij}(h), b_{ij}(h) = N) \\ &= \pi_j(a_{ij}(h), b_{ij}(h) = N) \\ &- \frac{max\, \pi_j(a_{ij}(h), N)|a_i \in A_i + min\, \pi_j(a_{ij}(h), N)|a_i \in E_i}{2}\end{aligned}$$

Hence, given that only $a_i = A$ is the only efficient strategy for player $i$ as discussed above, it follows that:

$$\kappa_i(a_{ij}(h) = A, b_{ij}(h) = N) = x_4 - x_4 = 0$$

$$\kappa_i\big(a_{ij}(h) = B, b_{ij}(h) = N\big) = x_6 - x_4 = -(x_4 - x_6)$$

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.*

# Chapter 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*, 24(9), 694-703.

**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.

**Baron, Marcia.** 1991. "Impartiality and Friendship." *Ethics*, 101(4), 836-57.

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

**Bauman, Christopher W.; A. Peter McGraw; Daniel M. Bartels and Caleb Warren.** 2014. "Revisiting External Validity: Concerns About Trolley Problems and Other Sacrificial Dilemmas in Moral Psychology." *Social and Personality Psychology Compass*, 8(9), 536-54.

**Becker, Lawrence C.** 1991. "Impartiality and Ethical Theory." *Ethics*, 101(4), 698-700.

**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.

**Blum, Lawrence.** 1991. "Moral Perception and Particularity." *Ethics*, 101(4), 701-25.

**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.

**Camerer, Colin F.; Anna Dreber; Felix Holzmeister; Teck-Hua Ho; Jürgen Huber; Magnus Johannesson; Michael Kirchler; Gideon Nave; Brian A. Nosek; Thomas Pfeiffer, et al.** 2018. "Evaluating the Replicability of Social Science Experiments in Nature and Science between 2010 and 2015." *Nature Human Behaviour*, 2(9), 637-44.

**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.** 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.

**Cookson, R.** 2000. "Framing Effects in Public Goods Experiments." *Experimental Economics*, 3(1), 55-79.

**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.

**Cottingham, John.** 1991. "The Ethics of Self-Concern." *Ethics*, 101(4), 798-817.

**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.

**____.** 2000. "Thinking Like a Game Theorist: Factors Affecting the Frequency of Equilibrium Play." *Journal of Economic Behavior & Organization*, 41(3), 299-314.

**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.

**Curry, Oliver S.; Darragh Hare; Cameron Hepburn; Dominic D. P. Johnson; Michael D. Buhrmester; Harvey Whitehouse and David W. Macdonald.** 2020. "Cooperative Conservation: Seven Ways to Save the World." *Conservation Science and Practice*, 2(1), e123.

**Curry, Oliver Scott; Mark Alfano; Mark J. Brandt and Christine Pelican.** 2021. "Moral Molecules: Morality as a Combinatorial System." *Review of Philosophy and Psychology*.

**Curry, Oliver Scott; Matthew Jones Chesters and Caspar J. Van Lissa.** 2019. "Mapping Morality with a Compass: Testing the Theory of 'Morality-as-Cooperation' with a New Questionnaire." *Journal of Research in Personality*, 78, 106-24.

**Cushman, Fiery.** 2013. "Action, Outcome, and Value:A Dual-System Framework for Morality." *Personality and Social Psychology Review*, 17(3), 273-92.

**____.** 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.

**Darley, John M. and Thomas R. Shultz.** 1990. "Moral Rules: Their Content and Acquisition." *Annual Review of Psychology*, 41, 525-56.

**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.

**Deigh, John.** 1991. "Impartiality: A Closing Note." *Ethics*, 101(4), 858-64.

**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.

**Ellemers, Naomi; Jojanneke van der Toorn; Yavor Paunov and Thed van Leeuwen.** 2019. "The Psychology of Morality: A Review and Analysis of Empirical Studies Published from 1940 through 2017." *Personality and Social Psychology Review*, 23(4), 332-66.

**Ellingsen, Tore; Magnus Johannesson; Johanna Mollerstrom and Sara Munkhammar.** 2012. "Social Framing Effects: Preferences or Beliefs?" *Games and Economic Behavior*, 76(1), 117-30.

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

**Everett, Jim A. C. and Guy Kahane.** 2020. "Switching Tracks? Towards a Multidimensional Model of Utilitarian Psychology." *Trends in Cognitive Sciences*, 24(2), 124-34.

**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.

**Feltham, Brian and John Cottingham.** 2010. *Partiality and Impartiality: Morality, Special Relationships, and the Wider World*. United States: Oxford University Press.

**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.

**Fischbacher, Urs; Simon Gächter and Simone Quercia.** 2012. "The Behavioral Validity of the Strategy Method in Public Good Experiments." *Journal of Economic Psychology*, 33(4), 897-913.

**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.

**Fleishman, John A.** 1988. "The Effects of Decision Framing and Others' Behavior on Cooperation in a Social Dilemma." *Journal of Conflict Resolution*, 32(1), 162-80.

**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.

**Friedman, Marilyn.** 1991. "The Practice of Partiality." *Ethics*, 101(4), 818-35.

**Friedman, Milton.** 1953. *Essays in Positive Economics*. Chicago, United States of America: The University of Chicago Press.

**Gächter, Simon.** 2014. "Human Prosocial Motivation and the Maintenance of Social Order," E. Zamir and D. Teichman, *The Oxford Handbook of Behavioral Economics and the Law.* New York, United States of America: Oxford University Press, 28-60.

**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.

**Gächter, Simon and Elke Renner.** 2010. "The Effects of (Incentivized) Belief Elicitation in Public Goods Experiments." *Experimental Economics*, 13(3), 364-77.

**Gellner, David N.; Oliver Scott Curry; Joanna Cook; Mark Alfano and Soumhya Venkatesan.** 2020. "Morality Is Fundamentally an Evolved Solution to Problems of Social Cooperation." *Journal of the Royal Anthropological Institute*, 26(2), 415-27.

**Graham, J.; Jonathan Haidt and B. A. Nosek.** 2009. "Liberals and Conservatives Rely on Different Sets of Moral Foundations." *Journal of Personality and Social Psychology*, 96(5), 1029-46.

**Graham, J.; B. A. Nosek; Jonathan Haidt; Ravi Iyer; Spassena Koleva and Peter H. Ditto.** 2011. "Mapping the Moral Domain." *Journal of Personality and Social Psychology*, 101(2), 366-85.

**Graham, Jesse.** 2014. "Morality Beyond the Lab." *Science*, 345(6202), 1242-42.

**Graham, Jesse and Jonathan Haidt.** 2010. "Beyond Beliefs: Religions Bind Individuals into Moral Communities." *Personality and Social Psychology Review*, 14(1), 140-50.

**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.** 2013. "Moral Psychology for the Twenty-First Century." *Journal of Moral Education*, 42(3), 281-97.

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

**____.** 2007. "The New Synthesis in Moral Psychology." *Science*, 316(5827), 998-1002.

**____.** 2012. *The Righteous Mind: Why Good People Are Divided by Politics and Religion*. New York, NY, US: Pantheon/Random House.

**Haidt, Jonathan and Jesse Graham.** 2009. "Planet of the Durkheimians, Where Community, Authority, and Sacredness Are Foundations of Morality," J. T. Jost, A. C. Kay and H. Thorisdottir, *Social and Psychological Bases of Ideology and System Justification.* New York, United States of Americs: Oxford University Press, 371-401.

**____.** 2007. "When Morality Opposes Justice: Conservatives Have Moral Intuitions That Liberals May Not Recognize." *Social Justice Research*, 20(1), 98-116.

**Haidt, Jonathan and Craig Joseph.** 2004. "Intuitive Ethics: How Innately Prepared Intuitions Generate Culturally Variable Virtues." *Daedalus*, 133(4), 55-66.

**____.** 2007. "The Moral Mind: How Five Sets of Innate Intuitions Guide the Development of Many Culture-Specific Virtues, and Perhaps Even Modules," P. Carruthers, S. Laurence and S. Stich, *The Innate Mind, Volume 3: Foundations and the Future.* New York: United States of America: Oxford University Press, 367-92.

**Haidt, Jonathan and Selin Kesebir.** 2010. "Morality," *Handbook of Social Psychology, Vol. 2, 5th Ed.* Hoboken, NJ, US: John Wiley & Sons, Inc., 797-832.

**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.

**Herman, Barbara.** 1991. "Agency, Attachment, and Difference." *Ethics*, 101(4), 775-97.

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

**Hintze, Jerry L. and Ray D. Nelson.** 1998. "Violin Plots: A Box Plot-Density Trace Synergism." *The American Statistician*, 52(2), 181-84.

**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.

**Hofmann, Wilhelm; Daniel C. Wisneski; Mark J. Brandt and Linda J. Skitka.** 2014. "Morality in Everyday Life." *Science*, 345(6202), 1340-43.

**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 States 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.

**Ioannidis, John P. A.** 2005. "Why Most Published Research Findings Are False." *PLOS Medicine*, 2(8), e124.

**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.

**Isler, Ozan; Simon Gächter; A. John Maule and Chris Starmer.** 2021. "Contextualised Strong Reciprocity Explains Selfless Cooperation Despite Selfish Intuitions and Weak Social Heuristics." *Scientific Reports*, 11(1), 13868.

**Iyer, Ravi; Spassena Koleva; Jesse Graham; Peter Ditto and Jonathan Haidt.** 2012. "Understanding Libertarian Morality: The Psychological Dispositions of Self-Identified Libertarians." *PLOS ONE*, 7(8), e42366.

**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.

**Kahane, Guy.** 2015. "Sidetracked by Trolleys: Why Sacrificial Moral Dilemmas Tell Us Little (or Nothing) About Utilitarian Judgment." *Social Neuroscience*, 10(5), 551-60.

**Kahane, Guy; Jim A. C. Everett; Brian D. Earp; Lucius Caviola; Nadira S. Faber; Molly J. Crockett and Julian Savulescu.** 2018. "Beyond Sacrificial Harm: A Two-Dimensional Model of Utilitarian Psychology." *Psychological Review*, 125(2), 131-64.

**Kahane, Guy; Jim A. C. Everett; Brian D. Earp; Miguel Farias and Julian Savulescu.** 2015. "'Utilitarian' Judgments in Sacrificial Moral Dilemmas Do Not Reflect Impartial Concern for the Greater Good." *Cognition*, 134, 193-209.

**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.

**Kocher, Martin G.; Todd Cherry; Stephan Kroll; Robert J. Netzer and Matthias Sutter.** 2008. "Conditional Cooperation on Three Continents." *Economics Letters*, 101(3), 175-78.

**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.

**López-Pérez, Raúl.** 2008. "Aversion to Norm-Breaking: A Model." *Games and Economic Behavior*, 64(1), 237-67.

**Malle, Bertram F.** 2021. "Moral Judgments." *Annual Review of Psychology*, 72(1), 293-318.

**Malle, Bertram F.; Steve Guglielmo and Andrew E. Monroe.** 2014. "A Theory of Blame." *Psychological Inquiry*, 25(2), 147-86.

**Marshall, Alfred.** 2013. *Principles of Economics*. Palgrave Macmillan.

**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.

**Mendus, Susan.** 2002. *Impartiality in Moral and Political Philosophy*. United States: Oxford University Press.

**Micallef, Luana and Peter Rodgers.** 2014. "Eulerape: Drawing Area-Proportional 3-Venn Diagrams Using Ellipses." *PLOS ONE*, 9(7), e101717.

**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.

**Piper, Adrian M. S.** 1991. "Impartiality, Compassion, and Modal Imagination." *Ethics*, 101(4), 726-57.

**Pizarro, David; Eric Uhlmann and Peter Salovey.** 2003. "Asymmetry in Judgments of Moral Blame and Praise:The Role of Perceived Metadesires." *Psychological Science*, 14(3), 267-72.

**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.

**Raphael, D. D.** 2009. *The Impartial Spectator: Adam Smith's Moral Philosophy*. United States: Oxford University Press.

**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.*

**Rutte, Christel G.; Henk A. M. Wilke and David M. Messick.** 1987. "The Effects of Framing Social Dilemmas as Give-Some or Take-Some Games." *British Journal of Social Psychology*, 26(2), 103-08.

**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.

**Schmitter, Amy M.** 2020. "Negociating Pluralism in Taste and Character: Reading the Second Enquiry with “of the Standard of Taste”," J. Taylor, *Reading Hume on the Principles of Morals.* New York, United States of America: Oxford University Press, 219-37.

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

**Selten, R.** 1967. "Die Strategiemethode Zur Erforschung Des Eingeschra¨Nkt Rationalen Verhaltens Im Rahmen Eines Oligopolexperimentes," H. E. Sauermann, *Beitra¨Ge Zur Experimentellen Wirtschaftsforschung.* Tübingen: J.C.B> Mohr (Paul Siebeck), 136-68.

**Sen, Amartya.** 1973. "Behaviour and the Concept of Preference." *Economica*, 40(159), 241-59.

**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.

**____.** 2017. "Sentiments, Conduct, and Trust in the Laboratory." *Social Philosophy and Policy*, 34(1), 25-55.

**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,

**van Dijk, Eric and Henk Wilke.** 1997. "Is It Mine or Is It Ours? Framing Property Rights and Decision Making in Social Dilemmas." *Organizational Behavior and Human Decision Processes*, 71(2), 195-209.

**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.

**Walker, Margaret Urban.** 1991. "Partial Consideration." *Ethics*, 101(4), 758-74.

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

**Wiltermuth, Scott S.; Benoît Monin and Rosalind M. Chow.** 2010. "The Orthogonality of Praise and Condemnation in Moral Judgment." *Social Psychological and Personality Science*, 1(4), 302-10.

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