

STEFANO DUCA

COOPERATION IN GROUPS:
A GAME-THEORETIC INVESTIGATION OF BEHAVIOUR,
MECHANISMS, AND DYNAMICS

DISS. ETH NO. 25753

COOPERATION IN GROUPS:
A GAME-THEORETIC INVESTIGATION OF
BEHAVIOUR, MECHANISMS, AND DYNAMICS

A dissertation submitted to attain the degree of

DOCTOR OF SCIENCES OF ETH ZURICH
(Dr. sc. ETH Zurich)

presented by

STEFANO DUCA
M.Sc.

born on 28 December 1988
citizen of Italy

accepted on the recommendation of

Prof. Dr. Dirk Helbing, examiner
Prof. Dr. Matjaž Perc, co-examiner
Prof. Dr. Rainer Hegselmann, co-examiner

2019

Stefano Duca: *Cooperation in Groups: a Game-Theoretic Investigation of Behaviour, Mechanisms, and Dynamics*, © 2019

DOI: 10.3929/ethz-a-

ABSTRACT

What makes people cooperate? How can one design mechanisms in order to incentivize players to contribute to public goods? These are the kinds of fundamental questions that are exemplified by analysis of the free-rider problem: The problem arises from the fact that, while an entire population benefits from the presence of a public good produced at some cost by cooperative individuals, free-riders (defectors) can benefit from the public good even when not producing any of it. Using the tools of Game Theory, Mechanism Design and Experimental Economics, one can identify and understand the underlying conflicting forces leading to such free-riders problems in human interactions. This understanding can then be used to design suitable mechanisms to avoid “tragedies of the commons”, i.e. convergence to socially sub-optimal outcomes.

In this dissertation, I focus on Public Goods games, and in particular on understanding under which conditions the public good is successfully provided and sustained through voluntary contributions, when players interact in groups. I not only focus on how cooperation can emerge as a result of incentive mechanisms and/or behavioural regularities, but I also study the implications of said mechanisms in terms of the total welfare of the players. The aim is to assess how robust positive predictions obtained for Voluntary Contribution Games are, when transferred to more general models which are closer to real-world social dilemmas situations.

I extend the Voluntary Contribution Game in several ways: by considering different strategy spaces and public good provision efficacies, by adding noise, and, crucially, by accounting for heterogeneity among players. Using a mixture of analytical, experimental and computational tools, I show that highly-efficient equilibria are enabled by so called “grouping” mechanisms but that they also often cease to exist when heterogeneity is taken into account. I identify under which conditions high cooperation can be achieved and determine what is the optimal mechanism in terms of social welfare, as a function of a social planner’s preference.

Finally, I also investigate mechanisms based on reputations expressed by “scores”, and show that the positive results obtained in pairwise interactions do not necessary apply to multiplayer prisoner’s dilemmas, regardless of how much information is provided about the past behaviour of the interacting partners.

ZUSAMMENFASSUNG

Was bewegt Menschen dazu sich kooperativ zu verhalten? Welche Mechanismen eignen sich dazu Menschen ausreichend zu motivieren, dass sie zum öffentlichen Gemeingut beitragen? Diese und mehr sind die Art von grundlegenden Fragen, die durch die Analyse "Trittbrettfahrer-Problems" veranschaulicht werden:

Das Problem entsteht dadurch, dass –obchon die gesamte Bevölkerung am Vorhandensein eines von kooperierenden Individuen produzierten All-gemeingutes interessiert ist– die "Trittbrettfahrer"(zu englisch «freerider») ebenfalls vom öffentlichem Gut profitieren können, ohne dass sie an der Produktion teil haben muessen. Indem man die Mittel und Methoden der Spieltheorie, der Mechanismus-Design-Theorie und der Experimentellen Ökonomik verwendet, kann man die zugrundeliegenden widerstreitende Kräfte in diesen Arten von Problemsituationen identifizieren und verstehen, und so die Dynamiken herausarbeiten die zu solchen Problemen mit "Trittbrettfahrern" menschlichen Interaktionen führen.

Dieses Wissen kann dann genutzt werden, um passende Mechanismen zu kreieren, und um die "Tragik der Allmende" zu vermeiden, d.h. die Konvergenz zu gesellschaftlich suboptimalen Kollektivergebnissen.

In dieser Dissertation konzentriere ich mich auf das so genannte «public goods game» (zu deutsch «Öffentliche-Güter-Spiel») und insbesondere darauf zu verstehen, unter welchen Bedingungen diese öffentlichen Güter in Gruppen erfolgreich bereitgestellt und durch freiwillige Mitwirkung erhalten werden.

Ich befasse mich hierbei nicht nur damit, wie Zusammenarbeit durch Anreizmechanismen und/oder Verhaltensregelmässigkeiten hervorgebracht werden kann, sondern ich untersuche vor allem auch die Auswirkungen gewisser Mechanismen mit Blick auf den resultierenden Gesamtwohland der Gruppe.

Mein Ziel ist es zu erörtern, wie solide positive Resultate gewisser Mechanismen sind, wenn sie auf generellere Modelle angewandt werden, die naeher an realen gesellschaftlichen Dilemmasituationen.

Ich erweitere das "Voluntary Contribution Game"(zu deutsch "Freiwillige Beitragsspiel") auf verschiedene Weise: durch die Berücksichtigung von Heterogenitaet und der diverser Synergiefaktoren, und durch das Hinzu-fügen von Imperfektionen des Mechanismus.

Unter Verwendung einer Kombination von analytischen, experimentellen und simulationsbasierten Methoden zeige ich, dass hocheffiziente Gleichgewichtszustände durch sogenannte "Gruppierungsmechanismen" ermöglicht werden, aber dass sie auch oftmals gänzlich wegfallen, wenn der Heterogenität Rechnung getragen wird.

Ich identifiziere unter welchen Bedingungen hochwertige Zusammenarbeit erreicht werden kann und bestimme den optimalen Mechanismus indem ich die Perspektive eines Sozialplaners einnehme.

Schliesslich untersuche ich die reputationsbasierten Mechanismen in Form von «Punktevergabe» und zeige auf, dass die positiven Ergebnisse, welche in paarweisen Interaktionen erzielt wurden, nicht notwendigerweise für Gefangendilemma mit mehr als zwei Spielern halten, egal wieviele Informationen über das vergangene Verhalten der interagierenden Spieler vorliegen.

RIASSUNTO

Quali sono le basi per la cooperazione fra persone? Come si possono progettare dei meccanismi che incentivino degli agenti a contribuire al bene pubblico? Queste domande fondamentali sono esemplificate dall'analisi del problema dei profittatori (*free-riders*): Il problema dei *free-riders* deriva dal fatto che ogni membro di una popolazione, anche chi non contribuisce al bene pubblico, beneficia della presenza dei beni pubblici prodotti dai cooperatori, che però ne pagano il costo di produzione. Usando gli strumenti della Teoria dei Giochi, del Mechanism Design e dell'Economia Sperimentale, è possibile identificare e comprendere le varie dinamiche contrastanti che fanno sì che molte interazioni umane diano luogo al problema dei *free-riders*. Una volta comprese le forze sottostanti a queste dinamiche, è possibile sviluppare dei meccanismi appropriati affinché si possa evitare che il sistema converga ad un equilibrio sub-optimale, la cosiddetta tragedia dei beni comuni (o *tragedy of the commons*).

Questa tesi è incentrata sui cosiddetti giochi dei Beni Pubblici focalizzandosi su quali siano le condizioni affinché il bene pubblico possa venire prodotto con contributi pienamente volontari quando i giocatori interagiscono in gruppi. Questa dissertazione studia come si possa far emergere un comportamento cooperativo sfruttando regolarità comportamentali e meccanismi in grado di incentivarne la diffusione. Essa analizza, inoltre, le implicazioni di tali meccanismi in termini di benessere collettivo della popolazione. In particolare, lo scopo di questo lavoro consiste nel valutare quanto robuste siano le predizioni positive (in termini di percentuale di cooperazione in una popolazione) ottenute per i cosiddetti *Voluntary Contribution Games*, quando estese a modelli più generici, e quindi capaci di descrivere i dilemmi sociali reali in maniera più realistica.

Questa tesi estende il modello dei *Voluntary Contribution Games* in diversi modi: espandendo lo spazio delle strategie possibili e l'efficacia della fornitura del bene pubblico, aggiungendo la possibilità di errori nelle osservazioni e, cosa fondamentale, tenendo conto dell'eterogeneità dei giocatori. Con metodi analitici, sperimentali e computazionali, è possibile dimostrare che si possono raggiungere equilibri molto efficienti, utilizzando meccanismi cosiddetti "raggruppanti" (cioè, meccanismi che sfruttano l'assegnazione dei giocatori in gruppi). D'altro canto, si dimostra che questi equilibri cessano di esistere quando si tiene in considerazione l'eterogeneità

tà dei giocatori. In questo lavoro vengono definite le condizioni affinché si possano raggiungere alti livelli di cooperazione e si descrive quale sia il meccanismo ottimale in termini di benessere collettivo, dal punto di vista di un pianificatore sociale.

Infine, questa tesi investiga vari meccanismi basati sulla reputazione appartenenti alla famiglia dei meccanismi di “scoring” e vuole dimostrare che i risultati positivi (in termini di cooperazione) che sono stati precedentemente ottenuti nel caso di interazioni tra due giocatori, non si applicano a dilemmi del prigioniero (o Prisoner’s dilemma) giocati da più di due individui. Nel caso di interazioni tra più persone, i risultati si confermano negativi a prescindere dalla quantità di informazioni riguardanti le azioni passate dei giocatori che viene fornita durante il gioco.

ACKNOWLEDGEMENTS

These years in Zurich have been an amazing experience, full of learning and personal growth and I have so many people to thank for this. First of all, I would really like to thank Dirk Helbing for his guidance, for always providing a stimulating environment, full of opportunities and academic freedom and for making sure that I enjoyed my time as a PhD student. My deepest gratitude goes to Heinrich, who accompanied my intellectual journey at ETH and taught me so much about so many topics, ranging from Game Theory to 80's pop culture. I truly appreciate his support and dedication (and patience sometimes). If there ever was an example of a collaborative effort, this dissertation embodies it! I would have never been able to complete this journey without the help and support of so many different people: First of all, my awesome colleagues at ETH: from my office-mates Kolja and Bary, to my fellow PhD students, Stefano, Thomas, Caleb, Farzam, Mark and Christian and the rest of COSS (including past members like Micheal Maes and Dietmar Huber). Thanks for always providing a fun and interesting environment, full of intellectual exchanges and discussions. A special shot-out goes to my coffee buddies for countless (and some might say endless) coffee breaks: thanks Stefano and Thomas for always being there to start them and to Caleb and John for making sure that at some point they would stop :-)) (seriously though, I don't know what I would have done without you). If these years in Zurich have been such a great and fun experience, it is also thanks to all my friends, old and new. Many thanks to Sandro, Silvia, John, Mari, Giuseppe, Gabbo, Roberta, Stefano, Kate, Per, Stuart, Petr and many others for being my home away from home and for providing moral support, limitless food and aperos, practical help, and very good company (Sandro and Silvia, I owe you so many home cooked dinners...). Also, thanks to Julia for the huge moral support while writing this thesis (and for "sometimes" listening to my complaints...) and for the help with everything German related. Many many thanks to all my friends all over the world, from Naples to Munich (go MCC!) and everywhere else for so many years of friendship and companionship, you guys are awesome! La mia gratitudine più profonda va alla mia famiglia: grazie a Zia Carmelina ed al resto della famiglia per tutto l'affetto e il supporto nel corso degli anni. Alle mie sorelle, Gabriella e Monica: grazie di tutto, per esserci sempre state e per avere stimolato la mia curiosità e voglia di

imparare, siete state il mio primo modello di riferimento e lo siete tuttora. Infine, un ringraziamento speciale va ai miei genitori, Paolo e Rosaria, per aver sempre creduto in me (anche quando oggettivamente era una cattiva idea) e per l'esempio di comportamento che mi hanno dato. Mamma e Papa', tutto ciò che ho fatto in questi anni e' stato possibile solo grazie alla vostra guida ed ai vostri sacrifici, grazie di tutto!

CONTENTS

1	INTRODUCTION	1
1.1	Preamble	1
1.2	Mechanisms	4
1.3	Behaviour	5
1.4	Dynamics	7
1.5	Manuscript contributions	10
1.6	Thesis overview	14
2	ASSORTATIVE MATCHING WITH INEQUALITY IN VOLUNTARY CONTRIBUTION GAMES	17
2.1	Introduction	18
2.2	The model	19
2.2.1	Simulation	20
2.3	Results	21
2.4	Summary of results	27
2.5	Appendix: Nash Equilibria	28
3	CONTRIBUTION-BASED GROUPING UNDER NOISE	35
3.1	Motivation	36
3.2	The Mechanism	37
3.2.1	The Model	37
3.3	Nash Equilibria	39
3.3.1	Game 1: 'Baseline'	40
3.3.2	Game 2: 'Heterogeneity Extension' (Extension 1)	41
3.3.3	Game 3: 'Noise Extension' (Extension 2)	42
3.3.4	Remark: Mixed-Strategy Nash Equilibria	43
3.4	Welfare Comparison	43
3.4.1	Homogeneous Endowments	44
3.4.2	Heterogeneous Endowments	46
3.5	Logit Dynamics	47
3.6	Summary	53
3.7	Appendix: Pure Strategy Nash Equilibria	55
3.7.1	Perfect Meritocracy	55
3.7.2	Fuzzy Mechanism	58
3.8	Appendix: Mixed-Strategy Nash Equilibria	61

3.9	Appendix: Properties of the Fuzzy Ranking	67
4	HETEROGENOUS AGENTS IN VOLUNTARY CONTRIBUTION GAMES WITH ASSORTATIVE MATCHING AND WEALTH ACCUMULATION	71
4.1	Introduction	72
4.2	The model	74
4.2.1	Ranking criteria	76
4.3	Results	77
4.4	Discussion	82
4.5	Methods	84
4.5.1	Efficiency and Gini coefficient	86
5	GROUPS AND SCORES: THE DECLINE OF COOPERATION	89
5.1	Introduction	90
6	CONCLUSION	109
A	APPENDIX: ASSORTATIVE MATCHING WITH INEQUALITY IN VOLUNTARY CONTRIBUTION GAMES	113
A.1	Efficiency loss as a function of the endowment distribution	113
B	APPENDIX: HETEROGENOUS AGENTS IN VOLUNTARY CONTRIBUTION GAMES WITH ASSORTATIVE MATCHING AND WEALTH ACCUMULATION	117
B.1	Simulation snapshots	117
B.2	Snapshots of wealth distribution	126
B.3	Average Cooperation Levels	130
C	APPENDIX: GROUPS AND SCORES: THE DECLINE OF COOPERATION	133
C.1	Detailed experimental results	133
C.2	Percentage of cooperators as a function of the observed score in their group	139
C.3	Further statistical analysis	146
C.4	Decision making micro model	151
C.4.1	Fit for first 12 rounds	155
C.4.2	Fit for all rounds	162
C.4.3	Marginal effects	168
C.4.4	Order Effects	173
C.5	Screenshots	174

BIBLIOGRAPHY 183

INTRODUCTION

The power of a theory is exactly proportional to the diversity of situations it can explain.

— Elinor Ostrom

1.1 PREAMBLE

According to Roger Myerson, Game Theory is the study of mathematical models of conflict and cooperation between intelligent rational decision makers [1]. First formulated by von Neumann and Morgenstern [2, 3], it provides general mathematical techniques for analysing systems in which two or more individuals make decisions that will influence one another's welfare.

To each combination of choices is assigned a payoff, i.e. a measure of how much each player gains when these strategies are selected. Generally (but not necessarily) it is assumed that players know the payoff assigned to each strategy and that they cannot communicate with each other; this implies that each player analyses the game and decides which strategy to play beforehand, hence players do not react to the actions of other players. A fundamental assumption is that each player is rational, i.e. that it always chooses to perform the action with the optimal expected outcome for itself¹. Under this assumption, each player can choose its optimal strategy, e.g. the strategy that leads to the highest payoff, by expecting all other players to choose their optimal strategy as well. A key result obtained by Nash [6] is that there always exist at least one choice of strategies for which no player can gain by unilaterally switching strategy; such a situation is called a Nash equilibrium.

Game Theory can be used to understand the conflicting forces that sum up to produce a very fundamental problem: the *Free Rider problem* [7]. The problem arises from the fact that, while an entire population benefits from the presence of a public good produced at some cost by cooperative individuals, free-riders (defectors) can still benefit from the public good with-

¹ Rationality imposes restrictions over the shape of the utility function, in particular about individuals' preferences over alternatives. A more detailed description of rationality is out of the scope of this dissertation; a curious reader can see e.g. [4, 5].

out producing any of it. In Game Theory terms, one can reformulate the problems as follows: A cooperator provides a benefit b to another player at a cost c with $b > c$; in contrast a defector does not provide any public good and thus it bears no cost. A defector, however, can still benefit from the public good produced by the cooperators. For rational agents, defecting is better than cooperating because it permits to avoid the cost of cooperation and the risk of being exploited, thus obtaining an advantage over cooperators. If all players defect, however, everybody obtains a net gain of 0, worse than the payoff $b - c > 0$ that results from diffused cooperation. Nevertheless, the only Nash equilibrium of this game is for all agents to defect and, as a result, they are all worse off than if everybody cooperated; a phenomenon often referred to as the “tragedy of the commons” [8]. Nonetheless, human societies offer many examples of people regularly working together (and thus cooperating) even when it would be in their selfish interest not to do so. In fact, some might say that the foundation of our modern societies is people working together to achieve a common good.

Situations in which individual strategic incentives result in under-production or over-consumption of goods and thus in sub-optimal outputs for the society, are often referred to as “social dilemmas”. Social dilemmas can be modelled by games sharing this common feature [9]: there exists at least one Pareto inefficient equilibrium, i.e. an equilibrium for which there exists an alternative outcome in which at least one player could be better off without making anyone else worse off. Games that display this inherent tension between the collective good (because a Pareto improvement could be possible) and the strategic incentives (which lead to the inefficient Nash equilibria) include Common Pool Resource management and Public Goods games².

The focus of this dissertation will be on Public Goods games. Common examples of public goods include: military defence, public infrastructures, clean air and other environmental goods, team work, scientific advances, Wikipedia. [12]. In particular, I will focus on a very fundamental question in economics and social sciences in general [13] (and indeed in several other disciplines): understanding the conditions under which public goods can be supplied through voluntary contributions.

The most common framework for modeling voluntary contribution Public Goods games, widely used by experimentalist and theoreticians, is the

² Common Pools and Public Goods are both examples of non-excludable scarce goods, the difference being that Common-pool resources are rivalrous while Public Goods are not. For a discussion on this topic please see e.g. [10, 11].

family of Voluntary Contribution Games (VCGs), first introduced by Marwell and Ames [14] and Isaac and Walker [15]. While there exist many variations of the VCG that take into account the fact that individuals might exhibit different enjoyment of public goods or that contributions might not be substitutable³ (for a literature review on VCGs from a modelling and experimental point of view, see e.g. [13, 18–20]), the most common instantiation of VCGs is the linear Voluntary Contribution Game as implemented by Andreoni [21]: N players make voluntary contributions choosing to contribute a percentage of their endowment (assumed to be equal for every player) and are then randomly assigned to equal-sized groups⁴. The payoff of each player is realized by aggregating all the contributions of the players in his group, multiplied by a marginal per capita rate of return (*mpcr* from now on), and then evenly dividing the result among the members of the group. Besides this, each player receives the part of his endowment which he did not contribute to the common pool. When the *mpcr* is bigger than the group size, the total welfare is maximized if every agent contributes all of its endowment. However, for any *mpcr* smaller than one, the only dominant strategy for each player is to contribute nothing. Hence, this game is a classic example of a social dilemma.

This dissertation focuses on Public Goods games with voluntary contributions. This introduction is meant to provide context for the contributions presented in the next chapters. In the next section, I briefly introduce the field of mechanism design and present a class of mechanisms that will feature prominently in this dissertation: Assortative Matching. The following section summarizes key insights on VCGs from a behavioural perspective and introduces the concepts of indirect reciprocity and reputation. Then, the dynamics section, introduces a prominent way to implement reputation dynamics, namely Image Scoring, and briefly discusses the concept of Logit dynamics. Readers already familiar with these concepts may directly go to the Manuscript Contributions section, where I summarize the results presented in this dissertation. A thesis overview concludes this chapter.

³ E.g. by a design with constant marginal value of the public good and randomly-varying, individual-specific values of the private good [16], or taking into account non linearities [17].

⁴ Alternatively, players can instead be assigned to a group in the beginning of the game and then groups would remain fixed throughout the game.

1.2 MECHANISMS 

In mechanism design one takes the point of view of a social planner designing the rules of a game in order to obtain or optimize a certain outcome⁵. It is often called reverse Game Theory because it starts by defining desirable outcomes and it work backwards to create a game that incentivizes players towards those outcomes. Hence, the incentives created by the choice of rules of games are central to the theory of mechanism design.

Mechanism design has classically been applied to a variety of problems such as auctions, optimal taxation systems, investments in public projects, voting, school choices and, more recently, distributed control [24–27]. Here I focus on mechanisms that can successfully incentivize players to contribute more to the public good. Indeed, the outcome of Voluntary Contribution Games dramatically depends on how the rules of the games are designed (for example on how players are assigned to the groups); it is therefore of interest to study mechanisms that can potentially overcome the Free Rider problem, thus solving the social dilemma.

A plethora of mechanisms has been suggested to maintain/explain large-scale cooperation among humans in Public Goods Games and Common Pool Resources management. A class of well studied and often applied mechanisms rely on governmental authorities to provide certain public goods to a population by levying taxes to pay for them [28]. Further, a government could employ subsidies (monetary or otherwise) in order to incentivize the production of a public good [29], an example would be providing tax breaks (the incentive) to citizens buying solar panel for their homes. A large body of scholarship has investigated the role of social norms [30–32], such as communal responsibility, and social ‘punishment’ [33], either by ‘peer punishment’ [34, 35] (where individuals decide to punish others in a bilateral way) or by ‘pool punishment’ [36, 37] (a tax-paid-like organization to which punishment is outsourced), on maintaining high levels of cooperative behaviour. However, this mechanisms tend to be invasive, and thus hard to implement in many settings. In this thesis, I instead focus on less intrusive mechanisms where, contrary to the ones mentioned above, no payoff additions, subtractions or transfer between individuals take place. One such family of mechanisms, that features prominently in my thesis, is “grouping” mechanisms.

⁵ Fudenberg [22] and Mas Colell [23] provide a good introduction to the field of Mechanism Design.

A prominent instantiation of these mechanisms has been proposed by Gunnthorsdottir et al. [38], who, instead of randomly matching players to group, suggest to group players based on each individual contribution: Players are ranked from best to worst based on how much they contributed, with ties among them broken at random, and are then assigned to groups accordingly. In particular, given a group size of S , the S players that contributed the most are put in the first group, the second S highest contributing players are put in the second group and so on. Gunnthorsondottir et al. prove that, if all are assigned the same initial endowment, there exists a Nash Equilibrium⁶ where almost all players contribute everything they have and a small amount of players contributes nothing. This equilibrium is almost maximally efficient in terms of total welfare, thus effectively solving the tragedy of the commons. Furthermore, it is important to note that the free-riding equilibrium continues to exist even in this new setting. A key property of the game, necessary for the nearly efficient equilibria to exist, is that ties in the ranking placement are broken at random. In all these equilibria, in fact, there exists a mixed group where fully contributive players are grouped with defectors, i.e. players contributing nothing to the common pool. In order for this to be a NE, the fully contributive players must have a sufficiently high probability to be grouped only with other full contributors so that they would not benefit (in expectation) by decreasing their contribution and be placed with certainty in the defectors group.

Interestingly, several experiments [39–41] have observed the existence of the high contribution equilibria when playing VCGs with Assortative Matching in the laboratory, and they have shown that, when existing, near-efficient equilibria are almost always preferred over worst ones.

1.3 BEHAVIOUR

In the following, I will briefly present some examples of behavioural mechanisms that have also been shown to be able to sustain high levels of cooperation. As opposite to mechanism design, where one tries to incentivize cooperation through cleverly manipulating the incentives structure, here one tries to exploit pre-existing behavioural patterns, in order to achieve the desired outcome, i.e. high levels of cooperation.

⁶ A Nash Equilibrium is an outcome of a game where no player has an incentive to deviate from his or her chosen strategy after considering an opponent's choice.

Indeed, when VCGs are played in laboratory experiments, it is often observed that people cooperate much more than it would have been predicted by the theoretical models (see e.g. [20, 42]).

This is often attributed to what has been deemed a “universal behavioural regularity” [43–45]: players are often *conditional cooperators*, i.e. they are willing to contribute more to a public good the more they observe other players contributing⁷. In particular, players often behave as “imperfect conditional cooperators” [46, 47], meaning that they contribute more when they observe more contributions, but they do so with a selfish bias in that they contribute less than the others do on average. This results in contributions “spiraling downwards” over time due to players observing less and less cooperative behavior thus further reducing their own contributions.

In 2-player games with fixed interactions (meaning that players always play with the same partner), conditional cooperation coincides with direct reciprocity. Direct reciprocity is the propensity for a person to cooperate provided that others cooperate with him/her as well, and with a view of what will happen in the future (also called ‘tit for tat’). It is important to note this could be part of a (proto-)strategic rule or of a repeated-game strategy.

Indeed, in general the norm of reciprocity is the expectation that people will respond favourably to each other by returning benefits for benefits, and maybe hostility with hostility [48]. Reciprocity as a social rule seems to come very natural to humans [49, 50]. As Gouldner wrote: “There is no duty more indispensable than that of returning kindness” [51]. Reciprocity has even been suggested to have played a fundamental role in the early development of human societies and markets [52, 53].

Let us picture a situation in which an individual has the opportunity to help another to gain a benefit for a certain cost, smaller than the gain of the other person. Under the assumption of direct reciprocity, a player would help the other person under the expectation that he would do the same, if the roles were switched. If the help was reciprocated in the next occasion, then both individuals would obtain a net benefit. If this game is repeated for several rounds, thus allowing for direct reciprocity, it can be shown that there can be many strategies that lead to both players always helping each other, depending on how forward-looking the players are [54, 55]. An example is the grim trigger strategy⁸ that prescribes cooperation (i.e.

⁷ Note that this could be a behaviour on its own, or be motivated by ‘other regarding’ preferences like inequality aversion or fairness.

⁸ See e.g. [56] for a discussion.

helping the other player) for the first round and then whichever strategy the other agent played the last round.

However, if two players were to interact only once, there would be no fear of retaliation (i.e. a switch to a defective strategy) and thus the mechanism of direct reciprocity could not work⁹. Just as direct reciprocity is based on repeated encounters between the same two individuals ("my behavior toward you depends on what you have done to me"), so is *indirect reciprocity* based on repeated encounters in a group of individuals ("my behavior toward you also depends on what you have done to others"). As David Haig said: "For direct reciprocity you need a face, for indirect reciprocity you need a name" [59].

In the context of indirect reciprocity, a person does not expect the recipient of his help to reciprocate but he expects that somebody else will. The driver of indirect reciprocity is often *reputation*, famously called by Milinski a 'universal currency for human social interactions' [60]. The idea is that a donor might provide help to somebody that is more likely to help others: Assuming that helping someone, or refusing to do so, can have an impact on a person standing within his community and that, when interacting, a donor will take into account the receiver's reputation to decide whether to help him or not, it follows that players would care about their reputation. Under these assumptions, it is 'rational' for 'forward looking' players¹⁰ to help others.

1.4 DYNAMICS

In this section I will briefly discuss which dynamics, based on the mechanism (behavioural and not) discussed above, can lead to sustained cooperation over a long period of time.

A simple way to implement a reputation dynamic driving indirect reciprocity is *image scoring*, first introduced by Nowak & Sigmund [61]. The idea is that each player has a score that represents its reputation among the other players. Every player starts with a score of 0. Whenever a player has the opportunity to help someone else, its score gets updated: if it helps the receiver its score is increased by one, if not it is decreased by one. Thus a player's reputation is continuously reassessed based on its last decision.

⁹ It has been pointed out that trigger strategies could still ensure a cooperative Nash equilibrium but that that would require some community-enforced mechanism [57, 58]

¹⁰ I.e. players that somehow take into account their future payoff when deciding about their course of action

Computer simulations have shown that the strategy to cooperate with anybody with a non-negative image score is the fittest one and it evolves to fixate in the population [61, 62]. This means that a population starting with agents playing multiple competing strategies, ranging from unconditional cooperation to unconditional defection, will eventually evolve in a population where the only existing strategy is conditional cooperation with anybody exhibiting a non-negative image score.

It is important to notice that in refusing to help an individual with a low image score, a player is decreasing his own score, thus reducing his own probability of receiving help in the future. In this way, not helping a player in bad standing can be interpreted as a form of punishment. Allowing the players to take into account also their own score, results in the best strategy being: helping another player if his image score is non-negative and the donor's own score is less than one [61, 62].

Besides image scoring, there are other possible strategies to implement indirect reciprocity. An example of a possible strategy that takes into account the aforementioned problem is the "standing strategy" introduced by Sugden [63]. In this model, everyone is initially in good standing and an individual loses good standing when refusing to help a receiver in good standing. It was found [64] that this strategy usually beats image scoring and can in fact invade a population of image scorers. Standing strategies are part of the "leading eight", the only eight rules that result in an evolutionary stable strategy with a high cooperation level as identified by Ohtsuki et al. in a study of all possible reputation dynamics¹¹ [65]. They conclude that keys to the success in indirect reciprocity are to be nice (maintenance of cooperation among themselves), retaliatory (detection of defectors, punishment, and justification of punishment), apologetic, and forgiving. It is important to note that image scoring is not part of the leading eight.

Numerous experiments have shown that donations are more frequent to receivers who had been generous to others in previous interactions, thus validating the concept of indirect reciprocity [66–69].

However, even though standing strategies possess superior analytical properties than image scoring, an experiment designed by Milinski et al. to explicitly distinguish experimentally between the two strategies obtained results compatible with image scoring but not with standing strategies [70]. It has been speculated [60] that a possible reason for why people seem to use image scoring in experimental settings is the simplicity of the rule.

¹¹ When reputation is considered binary: a player can have either good or bad reputation.

Standing strategies, indeed, require much more information to work than image scoring. To learn the standing of a player, it is not enough to know what his last action was but also what the last action of the last person with whom he interacted was and so on. In fact, it would be necessary to know the entire history of the game. With binary image scoring, instead, it is enough to know what the last action of the recipient was.

Therefore, another aspect to consider is how much information each reputation rule requires, because many real world situations take place in settings that don't allow for players to observe the actions undertaken by all the other individuals. For instance, if agents are playing a public good games in groups, information regarding the individual behaviour can be hard to obtain. The performance of the group as a whole could be the only available information. It was shown [71] that in group interactions, when only the group performance is available but not the individual information about the players, cooperation cannot be sustained. As a measure of group performance, it is necessary to introduce the concept of group score. Each player's group score summarizes the aggregate cooperativeness of the groups of which he has been a member in the past, without any additional information regarding what the player individually did. It is also possible to show that, if players are given with probability p the image score of the players with which they interact and with probability $1 - p$ the group score, a low value of p is enough to lead to cooperation.

It thus seems that scoring can result in the emergence of cooperation for a wide range of informational settings.

In the discussion above, I have approached the dynamics of the games from a mainly behavioural point of view. In the following, I will briefly discuss an approach that is particularly suited to implement game dynamics in computer simulations. Many possible dynamical rules that could be used to simulate the outcome of a game have been suggested in the literature¹²; in this dissertation, I focus on best response dynamics [76], and in particular on perturbed best response [77, 78]: The family of (myopic) best response dynamics implements strategy updating rules where players decisions are determined by maximizing their payoff in the current round (potentially with mistakes/noise) via best responding to the actions taken by all the other players in the previous round. Perturbed best response expands this concept to a probabilistic choice when updating one's strategy; in this way, it is ensured that the probability of an agent's choice varies

¹² Examples of such rules are, among others, imitation [72], replicator dynamics [73], regret minimization [74] and learning [75].

smoothly with changes in the payoff values, thus obtaining differentiable dynamics [79, 80]. A crucial feature of perturbed best response is that the probability of choosing a certain strategy is proportional to the payoff that such strategy would realize, thus maintaining the concept of a rational agent.

In particular, in this dissertation I make extensive use of the Logit dynamic [81]¹³, which offers a best response behavioral model built on the concept of bounded rationality agents, i.e. agents whose rationality is somehow limited (e.g., due to their cognitive limitations or the tractability of the decision problem) [84]. The Logit dynamics allows to obtain a smooth approximation of the myopic best response dynamics while being very suited to a computational approach [85].

1.5 MANUSCRIPT CONTRIBUTIONS

The mechanisms discussed above would let societies overcome the free riders problem without the need of additional structures. Crucially, however, the positive predictions arising from these mechanism rely on specific underlying assumptions that might not be satisfied in more realistic scenarios.

In this dissertation, using a mixture of analytical, computational and experimental tools, I set out to address the important question of how robust the high-contributions predictions are, under more relaxed and realistic assumptions. The focus will be on voluntary contribution games where individuals interact in groups, which can be formed on the basis on various criteria.

As a first step toward assessing the robustness of the positive equilibria resulting from assortative matching, I examine when do the nearly efficient equilibria exist for a wider range of public-goods provision efficacies (that nests the standard *mPCR* model as a special, linear case) and for different action spaces. This can be relevant when observing these classes of games empirically: it could be quite hard to exactly determine the payoff structure of the game and/or the players' available actions, and thus it is important to know how robust a theoretical prediction is when the structure of the game is changed slightly. I find that the equilibria predictions are not dependent on the exact nature of the strategy space of the game but that they are indeed dependent on "how good" the public-good provision efficacy

¹³ The Logit is related to the concept of Quantal Response Equilibrium. For a discussion on the topic, see e.g. [82, 83].

is: Counter-intuitively, when the public-goods provision efficacy is so good that the nearly efficient outcome becomes highly rewarding, the only existing Nash equilibrium is for all players to contribute almost nothing to the common pool, thus disrupting the positive predictions obtained by Gunthorsdottir et al. . This is due to the fact that, while contributing would yield a huge benefit for the group, defecting becomes tempting for too many players, thus making it impossible for a mixed group of contributive and defective players to exist; a condition necessary for the highly efficient equilibria to exist, as discussed in section 1.2.

A focal point of this dissertation is to tackle a crucial limitation of previous studies: allowing for heterogeneous players. Undeniably, players are often different from each other: e.g. workers might have different innate talent no matter how hard they work, investors might have different abilities to judge a good investment or start with a different initial capital. Furthermore, there is evidence in the literature that taking into account the heterogeneity of players' attributes or endowments might result in different outcomes than in the homogeneous case, for several different games. For example, Cherry et al. [86] show that in a linear public good game players tend to contribute much less if provided with heterogeneous endowments¹⁴. However this key characteristic has been missing in the models described above. In this dissertation, I take into account two diverse dimensions over which agents can differ: wealth and talent.

Using analytical tools, I show that the consequences of including heterogeneous agents in the model depend crucially on the structure of the game, but that they can be quite disrupting. Indeed, when players pick their strategy from a continuous action space, all near-efficient Nash equilibria that exist under homogeneity fall apart. Instead, either players revert back to the negative full-defection equilibrium or new, previously impossible, complex mixed-strategy Nash equilibria emerge. In fact, when players have access to a continuous action space, they can reduce or increase their contributions by an infinitesimal amount. Due to the players' heterogeneity, it will always be beneficial to do so (see chapter 2 for details) and thus there cannot exist an equilibrium where two or more players contribute the same amount, a condition for the high-efficient equilibria to exist. Using computational tools, it is possible to determine the loss in efficiency of the mixed strategy equilibria, compared to the homogeneous case.

¹⁴ For other examples see [87, 88].

In the case of a binary¹⁵ or very coarse action space, instead, it is possible to prove that the highly efficient equilibria continue to exist. This is because in this case, the above argument about infinitesimal changes in contributions does not hold anymore.

This manuscript also addresses the case of imperfect grouping mechanisms. The imperfection is realized by adding a certain amount of noise in the ranking procedure (or equivalently by adding noise to the observation of each player's contribution). Hence, in case of full or no noise, the model reduces to the ones described above. It is possible to prove that, if the noise in the ranking is not too high, whenever the nearly efficient equilibria exist in the case of perfect matching, they can exist also for the imperfect mechanism (albeit for a higher *mpcr*). By interpreting the noise as a policy instrument, I assess the welfare properties of the different equilibria: it is found that in some case a small amount of noise in the matching mechanism can be beneficial, allowing for a fairer distribution of wealth while conserving the population's efficiency.

Building on these results, I address another important limitation of prior studies: namely the focus on one-shot games¹⁶ instead of the more realistic repeated interactions. Moving from one to repeated interactions can substantially alter the equilibria of the game especially if one allows for wealth to be accumulated over time; for one, because it might increase the heterogeneity of the population over time which in turn could alter the level of cooperative behaviour. With the introduction of different sources of agents' diversity, one has automatically introduced different criteria based on which one could assign agents to groups. For example, one could rank the players based on the total contribution that they made to the group or based solely on the percentage of their endowment they contributed, regardless of their talent (i.e. a multiplicative factor in front of the *mpcr*, making it potentially unique for each player).

Using an agent based computer simulation where agents follow a logit-response dynamics, I find that how much agents cooperate dramatically depends on which criteria it is used to rank the agents. Moreover, criteria that result in highly cooperative behaviour on the short run, might fail to sustain it over a longer time span. The distribution of wealth among the population is also a function of which criterion is used to assign agents to groups. Treating the choice of the ranking criterion as a policy choice, I find that a social planner trying to minimize the (wealth) inequality in

15 Meaning that the only possible actions are to contribute everything or nothing to the common pool.

16 Meaning a game where players interact only once.

the population while maximizing society's output would prefer ranking agents only based on their relative contribution to the common pool if very inequality averse, and ranking agents taking into account their talent, but not their total amount of wealth otherwise.

Repeated interactions among players are also the focus of the final contribution of this dissertation. As discussed in section 1.3, for forward-looking enough players, image scoring is a known mechanism that can potentially lead to the emergence of cooperative behaviour through indirect reciprocity. Indeed, many laboratory experiments have found results that confirm this in two-players interactions. However, in reality most social interactions unfold in groups and yet no laboratory experiment in a group setting has been performed so far. Extending scoring mechanisms to group interactions, presents interesting challenges: real-world group interactions vary with respect to the information that is available, and typically individuals do not observe all actions undertaken by all other individuals, especially in large groups. Moreover, the larger the groups, the more difficult it might be to correctly process all the available information. Hence, when players interact in a group setting, there are many potential scoring rules that could be applied.

Another appealing question is whether a centralized authority is needed for scoring to work. For the scores to be meaningful, in fact, there has to exist some sort of mechanism that keeps track of the individual contributions and then assigns the scores based on the performance of the players/groups. A centralized authority could perform this role by observing the actions of the players and then assigning the scores. However, there could be situations where it would be impossible to simultaneously observe the actions of all the players, e.g. due to logistic reasons. Furthermore, even if such a mechanism could be put in place, there might be concerns regarding privacy or the misuse of the collected information [89]. Hence, it is interesting to investigate a decentralized mechanism where players themselves can rate their fellow group members based on their contributions in the last round. Naturally, the decision to assign a certain score to a player could, of course, be subjected to strategic consideration, and thus it is an interesting question to ask how would the players rate their group-mates.

Hence, the goal is to find out how much information is enough for cooperation to be sustained in a group setting, considering various informational contexts. For this, a laboratory experiment was designed, focusing in particular on the simplest possible implementation of scoring mechanisms: 'Markovian' scores, i.e., scores that depend only on the players' ac-

tions from the previous period¹⁷. The baseline of the experiment was to test image scoring in a group setting. In addition, alternative scoring rules that could apply to group interactions were tested, including one where players score each other endogenously through votes. The proposed rules differ with respect to how much information regarding past behaviour of their group-mates is required, ranging from no feedback to full feedback.

The experimental results concerning cooperation are unambiguously negative: for every scoring mechanism, a steady decline in cooperation is observed. The decay of cooperation is similar under every mechanism and comparable even with the case when no scoring mechanism at all is implemented. This applies even for the scoring mechanism discussed in section 1.3 that was experimentally shown to stabilize high level of cooperative behaviour in the two-player case. A plausible explanation for this, is that it is harder to isolate the 'bad apples' in a group interaction; i.e. defectors cannot be individually punished, and cooperators cannot be individually rewarded. This results in a reaction to the average group score that is increasingly biased towards defection, therefore leading to a steady decrease of high-reputation players in the population that in turn begets lower levels of cooperative behaviour. Furthermore, in the cases where only group-level information is provided to the players, when a high-score player decides not to cooperate because of the presence of low-score subjects in his group, this reduces the score of all his group-mates, not just of the low-score individuals. This results in a steady decay of players with good reputation and cooperative behaviour in the population, and consequentially to a downward spiral of contributions, akin to the one observed by Fischbacher et al. and discussed above.

1.6 THESIS OVERVIEW

The overarching research question of this thesis is: what are mechanisms that can support high level of cooperative behaviour over a long period of time and, crucially, how robust/reliable are they? All material contained in this cumulative thesis is based on scientific publications addressing these questions. Each chapter of the dissertation is based on an individual paper: Three of these papers are already published in peer-reviewed international journals, while the other one is currently under review in a peer-reviewed

¹⁷ This is akin to assume that players only value/remember their last interactions with other players, i.e. that players have memory 1.

international journal. A list of these papers is provided at the end of this dissertation.

The rest of the manuscript is constructed as follows:




Chapter 2 is based on a paper titled: **Assortative Matching with Inequality in Voluntary Contribution Games** Using a mixture of analytical and computational tools, it analyzes how robust the positive equilibria predictions of VCGs with Assortative Matching are, when a wider range of public-goods provision efficacies is considered and agents have heterogeneous endowments.

Chapter 3 is based on a paper titled: **Contribution-Based Grouping under Noise**. Using a mixture of analytical and computational tools, it extends the above generalization to a wider set of strategy spaces while at the same time allowing for inaccurate (noisy) implementations of the group-matching mechanism. Furthermore, interpreting the noise as a policy instrument, it investigates the welfare properties of the assortative matching mechanism.

Chapter 4 is based on a paper titled: **Heterogenous agents in voluntary contribution games with assortative matching and wealth accumulation**. Using an agent based computer simulation, it builds on the previous chapters to analyze the long term outcome of repeated VCGs with Assortative Matching under different group-matching criteria, accounting for the possible heterogeneity in wealth and talents of the agents and allowing agents to accumulate wealth over time.

Chapter 5 is based on a paper titled: **Groups and scores: the decline of cooperation**. Using data from a behavioral laboratory experiment, it proposes and tests several scoring rules (implementing a reputation dynamic as a driver of indirect reciprocity) that could apply to VCGs, played in separate groups. In addition, it tests an alternative scoring rules where players score each other endogenously through votes.

A discussion concludes and an Appendix contains the Supplementary Materials of the papers presented above.

Even though all of the above chapters touch upon all the different facets of this dissertation, they each focus on one or more of these aspects. In the table below, I provide a comparison of the chapters of this thesis, indicating their focus, on which paper they are based and what is their publication status. The focus of each paper is represented by one or more icons: the gears () indicate a focus on mechanisms, the dynamical lines () a focus on the dynamics, and the three people () a focus on behaviour.





Chapter	Focus	Paper	Journal
2		Assortative Matching with Inequality in Voluntary Contribution Games	<i>Computational Economics</i>
3		Contribution-Based Grouping under Noise	<i>Games</i>
4		Heterogenous agents in voluntary contribution games with assortative matching and wealth accumulation	Under review in <i>Palgrave Communications</i>
5		Groups and scores: the decline of cooperation	<i>J. R. Soc. Interface</i>

TABLE 1.1: Comparison of the chapters in this PhD Thesis.

ASSORTATIVE MATCHING WITH INEQUALITY IN VOLUNTARY CONTRIBUTION GAMES

ABSTRACT

Voluntary contribution games are a classic social dilemma in which the individually dominant strategies result in a poor performance of the population. The negative zero-contribution predictions from social dilemma situations give way to more positive (near-)efficient ones when assortativity, instead of random mixing, governs the matching process in the population. Under assortative matching, agents contribute more than what would otherwise be strategically rational in order to be matched with others doing likewise. An open question has been the robustness of such predictions in terms of provisioning of the public good when heterogeneity in budgets amongst individuals is allowed. Here, we show analytically that the consequences of permitting heterogeneity depend crucially on the exact nature of the underlying public-good provision efficacy, but generally are rather devastating. Using computational methods, we quantify the loss resulting from heterogeneity vis-a-vis the homogeneous case as a function of (i) the public-good provision efficacy and (ii) the population inequality.

2.1 INTRODUCTION

Suppose a population of agents faces the *collective action* [90] challenge to provide public goods by means of various simultaneous but separate *voluntary contributions games* [91]. In each one, the collective would benefit from high contributions but individuals may have strategic incentives [6] to contribute less. Such situations, also known as ‘social dilemmas’ related to collective management of ‘common-pool resources’ [30, 92], often result in underprovisioning of the public good (i.e. tragedy of the commons as in [8]) as a result of this misalignment of collective interests and strategic incentives.

Generally, grave underprovision of the public good is the unique Nash equilibrium when individual contribution decisions are independent of the matching process. [21]’s model with random re-matching of groups is the best-known instantiation of this, and numerous experimental investigations have reported corresponding decays in contributions [20, 93] when such games are played in the laboratory [20, 93]. Predictions may change dramatically, however, when agents are matched ‘assortatively’ instead, that is, based on their pre-committed choice on how much to contribute so that high (low) contributors are matched with other high (low) contributors. Such mechanisms have been coined ‘meritocratic group-based matching’ [38], short ‘meritocratic matching’ [94].¹ Under meritocratic matching, new equilibria emerge through assortative matching that are as good as near-efficient [38, 94]. Indeed, when better (i.e. more efficient) equilibria exist, humans have been shown to consistently play them in controlled laboratory environments [38–41].

In this chapter, we address the important question of how robust the positive predictions stemming from assortative matching are. To assess this, we generalize the baseline model on two dimensions. On the one hand, we consider a range of public-goods provision efficacies that nests the standard marginal-per-capita-rate-of-return (‘mpcr’) model as a special, linear case. On the other hand, we allow heterogeneity in players’ budgets, expressing the *ex ante* inequality amongst individuals. In other contexts, heterogeneity has been shown to ‘help’ cooperation [96]. Our work, in particular, builds on one prior attempt at generalizing the standard model in terms of heterogeneity by [39], who consider two levels of budgets in the standard case of mpcr-linear payoffs.

¹ Note that this kind of mechanism differs crucially from other contribution-inducing mechanisms such as ‘punishment’ [95, 96] as no payoff additions, subtractions or transfer between individuals take place.

Methodologically, we blend analytical and computational approaches. Our results summarize as follows. We show analytically that the consequences of permitting heterogeneity in terms of provision of the public good depend crucially on the exact nature of the underlying public-good provision efficacy, but generally are devastating. Indeed, all near-efficient Nash equilibria that exist under homogeneity fall apart when heterogeneity is allowed. Instead, we are either back at the negative all-contribute-nothing equilibrium or new, previously impossible, complex mixed-strategy Nash equilibria emerge. In the latter case, the expected level of resulting public-good provision depends crucially on (i) the public-good provision efficacy and (ii) the population inequality. These mixed equilibria are virtually impossible to characterize and to evaluate analytically for general cases. We therefore use computational methods and quantify the loss resulting from heterogeneity vis-a-vis the homogeneous case as a function of parameters regarding (i) and (ii). Our analysis thus permits precise statements regarding possible consequences in terms of making wrong predictions when assuming a homogeneous population, which in many real-world cases is unrealistic.

The rest of this chapter is structured as follows. Next, we set up the model including details about our computational algorithm. Section 3 contains the chapter's results. Section 4 concludes. Appendix 2.5 contains details of the analytical results.

2.2 THE MODEL

N players are assigned an initial endowment w_i , that might be different for each player i , and play the following game, of which all aspects are common knowledge:

1. *Actions.* Each player makes a simultaneous and committed unilateral decision regarding how much of his endowment to contribute. We will indicate with $\alpha_i \in [0, 1]$ the percentage of w_i contributed by player i . We indicate with $\alpha = \{\alpha_i\}$ the array of strategies obtained in this way and with α_{-i} the strategy vector obtained excluding agent i .
2. *Matching.* Players are ranked by their effective contributions $s_i = w_i \alpha_i$ (from highest to lowest with random tie-breaking). They are then assigned to $M = \frac{N}{S}$ equal-sized groups of size S , such that the S highest-ranking players are assigned to the first group, the S second-highest ranking players are assigned to the second, etc.

3. *Outcome.* Payoffs ϕ_i realize based on the contribution total in each group. Each player receives the amount that he did not contribute plus the sum of all the contributions made by the members of his group multiplied by a factor Q (the marginal-per-capita-rate-of-return):

$$\phi_i(\alpha_i \mid \alpha_{-i}) = w_i(1 - \alpha_i) + Q \sum_{j \in G_i} \alpha_j^\gamma w_j \quad (2.1)$$

with G_i being the group to which agent i belongs.

The mpcr Q represents the benefits of cooperation among members of the same group. When $\frac{1}{5} < Q < 1$ the game is a social dilemma, meaning that group efficiency is maximized if every member contributes fully but doing so is always a dominated strategy.

The parameter γ in the return from the common pool is a measure of the “goodness” of the public-good provision efficacy. For high values of γ (super-linear payoffs) even a high percentage of contributions has little effect on the common pool, thus making the public good less fruitful. On the other end, for low γ values (sub-linear payoffs) even a small contribution allows players to obtain large benefits from cooperation. It is important to notice that the value of γ determines also the maximum efficiency² of the system ranging from $w_i + Q \sum_{j \in G_i} w_j$ for $\gamma \rightarrow 0$ to w_i for $\gamma \rightarrow \infty$.

2.2.1 Simulation

We simulate the above model in the following way:

1. We assign the initial endowments to each player sampling them from a truncated Gaussian distribution with mean W_0 and standard deviation σ . The distribution is truncated at $w_i = 0$ and $w_i = 2W_0$ so that endowments are always positive and symmetrically distributed around W_0 .
2. The initial strategy α_i of each player is set to 0, hence the simulation starts in the fully defective state³.

2 Efficiency is defined as the gain from the game relative to the initial endowment:

$$Efficiency = \frac{\sum_{i=1}^N \phi_i - \sum_{i=1}^N w_i}{\sum_{i=1}^N w_i}$$

3 Different initial starting conditions have been explored and they have been observed to have no effect on the final outcome of the simulation.

3. Each player updates his strategy in the following way:
 - With probability p he keeps playing the strategy played the round before.⁴
 - With probability $1 - p$ he myopically best responds to the strategies played by the other agents the round before: The agent checks what would his rank and consequent payoff be for each of the possible strategies⁵ he can play, given that the other agents keep playing the strategies played the round before. He then chooses the strategy that results in the highest payoff.
4. Based on the contribution of each player, groups are formed as described above and payoffs are then materialized.
5. The endowment of each player is then reset to his initial one and the algorithm repeats from point 3.

After T rounds the algorithm stops and the population average of the strategies is computed. The procedure is repeated EN times and the ensemble average of the population average is obtained.

2.3 RESULTS

The above game exhibits different Nash equilibria depending on the value of γ and on the players having different or the same initial endowments.

As already shown in [38], in the case of linear payoff ($\gamma = 1$) and homogeneous players the voluntary contribution game with assortative matching has multiple pure strategy Nash Equilibria: one is non contribution by all players and the others are almost Pareto optimal equilibria in which nearly all players contribute their entire endowment and few (less than the group size) contribute nothing. It is easy to see (see Appendix 2.5) that this result actually holds for any value of γ bigger than a threshold value $\bar{\gamma}$, with $\bar{\gamma} < 1$ and depending on groups size, mpcr and the total number of players. The only difference with the linear payoff case is that for $\gamma < 1$ the within group Nash Equilibrium is to contribute $\bar{a} = \left(\frac{1}{Q\gamma}\right)^{\frac{1}{\gamma-1}}$. Hence, for sublinear payoffs such that $\gamma > \bar{\gamma}$, the Nash Equilibria are: all players contribute \bar{a} and all players contribute their total endowment except for

⁴ Inertia was added to ensure the convergence to the pure-strategy Nash equilibrium (if existing). Without inertia, the best-response dynamics could oscillate around such an equilibrium.

⁵ The interval $[0, 1]$ is, of course, discretized.

few players that contribute $\bar{\alpha}$. For very small values of the exponent of the public good provision, the numbers of non contributors in the near efficient equilibrium becomes too high to be sustained. Hence, even though the public good game would be very convenient, for $\gamma \leq \bar{\gamma}$ the only existing equilibrium is that all players contribute $\bar{\alpha}$ ⁶.

A key property of the game, necessary for the nearly efficient equilibria to exist, is that ties in the ranking placement are broken at random. In all these equilibria, in fact, there exists a mixed group where fully contributive players are grouped with defectors, i.e. players contributing the within group Nash Equilibrium. In order for this to be a NE, the fully contributive players must have a sufficiently high probability to be grouped only with other full contributors so that they would not benefit (in expectation) by decreasing their contribution and be placed with certainty in the defectors group.

If each player is endowed with a different initial wealth, however, things change drastically. The heterogeneity of the players implies that there cannot be an equilibrium where more than one player contributes the same positive percentage of his endowment: Since players are ranked based on their effective contributions $s_i = w_i \alpha_i$, if more than one player were to contribute the same percentage, the one with the highest endowment would have a profitable deviation due to the existence of the mixed group. He could in fact contribute slightly more and be guaranteed to be placed in a better group. If the two players were already contributing everything, the wealthiest player could instead contribute slightly less and still be guaranteed to remain in the same group. For the above reason, the nearly efficient Nash Equilibria in which almost all players contribute everything does not exist for heterogeneous players. Moreover, any unique contribution α_i such that $\bar{\alpha} < \alpha_i < 1$ is also clearly not a Nash Equilibrium due to the fact that it would be possible to contribute less and still be placed in the same group.

Furthermore, for sublinear payoffs ($\gamma < 1$) the equilibrium in which all players contribute the within group Nash Equilibrium does not exist. Indeed, for $\gamma < 1$ we have that $\bar{\alpha} > 0$ and if all players were to play $\bar{\alpha}$, players with a lower endowment would have a profitable deviation by increasing their contributions and being grouped with players with a higher endow-

6 All players contributing $\bar{\alpha}$ is always an equilibrium for homogeneous players. The proof proceeds like in the linear case: it does not make sense to be the only player contributing more than that because this would only make the player's groupmates better off at the player's expense. Contributing less than $\bar{\alpha}$ is never beneficial, not even with random re-matching of groups, because of the non-linearity of the payoff function. See Appendix A for more details.

	$\gamma \leq \bar{\gamma}$			$\bar{\gamma} < \gamma < 1$			$\gamma \geq 1$		
	PSL	MS	PSH	PSL	MS	PSH	PSL	MS	PSH
Homog.	✓	✗	✗	✓	✗	✓	✓	✗	✓
Heterog.	✗	✓	✗	✗	✓	✗	✓	✗	✗

TABLE 2.1: In this table we show which equilibria exist for homogeneous and heterogeneous players depending on the value of γ . For homogeneous players, for any payoff such that γ is bigger than a threshold value $\bar{\gamma}$, there exist one Nash Equilibrium in which all players contribute nothing (indicated as PSL) and almost Pareto optimal Nash Equilibria where almost all players contribute everything and few (less than the group size) contribute nothing (PSH). For $\gamma < \bar{\gamma}$, the only Nash Equilibrium is non contribution by all. In these cases, there exist no mixed strategy (MS) Nash Equilibrium. For heterogeneous players the situation is different for different values of γ . For sublinear payoffs ($\gamma < 1$) there exist no pure strategy equilibria and hence the only Nash Equilibrium is in mixed strategies. For superlinear payoffs ($\gamma \geq 1$), the only pure strategy Nash Equilibrium is non-contribution by all players.

ment. Hence, for heterogeneous players and sublinear payoffs there exist no pure strategy Nash Equilibria for the game.

For superlinear payoffs ($\gamma \geq 1$), however, the within group Nash Equilibrium is to contribute nothing and hence the pure strategy equilibrium in which no player contributes anything continues to exist. Consequently, there are no mixed strategy Nash Equilibria for this game.

Table 2.1 summarizes which Nash Equilibria exist in which situation. In Appendix 2.5 we formally derive the results described in this section.

To obtain the mixed strategy equilibrium of the game we resort to computational methods. We simulate agents playing the game for different payoffs and different width of the initial wealth distribution.

We are mainly interested in a comparison between the (unique) equilibria in the case of heterogeneous players and the equilibria reached in

the homogeneous case. In particular, we are interested in how much efficiency² is lost due to the heterogeneity of players. Indeed, even though non contribution by all is a Nash Equilibrium, the quasi Pareto optimal equilibrium is payoff dominant⁷ and hence is the one to which we refer (when it exists). Furthermore, experimental results [38] have also shown that the nearly efficient equilibria are the one reached by the population.

In figure 2.1 we plot the loss of efficiency due to the heterogeneity of the players with respect to the level of contributions that would have been achieved in the homogeneous case as a function of γ and for a choice of representative parameters. A 100% loss (dashed red line) indicates that all players contribute nothing and thus that there is a complete loss of efficiency with respect to the homogeneous case. A 0% loss (dashed green line) means that the system reaches the same efficiency as it would with homogeneous players. A negative loss indicates that when endowments are heterogeneous the equilibrium reaches a higher efficiency than in the homogeneous case.

We observe, as predicted, that for superlinear payoffs the only possible equilibrium is non contribution by all and thus that the loss of efficiency is total. For intermediate sublinear payoff the system achieves different efficiency level, from quite low ones to levels closer to the homogeneous case. The quantitative value of efficiency that the mixed equilibrium achieves depends on the value of the mpcr and the width of the distribution of initial wealth as well as from other parameters. For values of the exponent of the public good provision lower than $\bar{\gamma}$, we initially observe an increase in efficiency when endowments are heterogeneous. This is due to the fact that for values of γ slightly below $\bar{\gamma}$, the only equilibrium in the homogeneous case is to contribute \bar{a} but it would still be more efficient if more players contributed a higher percentage of their endowment. Finally, for $\gamma \rightarrow 0$, the heterogeneous system approaches the same efficiency of the homogeneous, on account of the benefits of cooperation being obtainable for an arbitrary small contribution.

Interestingly, one can observe that the efficiency loss doesn't seem to change much for different widths of the endowment distribution (see figure A.2 in the appendix). For the width of the distribution approaching 0 we observe, as expected, the Nash Equilibria in case of homogeneous endowments.

⁷ Here we use payoff dominant in the sense of [97, 98]. The nearly efficient equilibrium can be shown to be ex-ante payoff dominant [38].

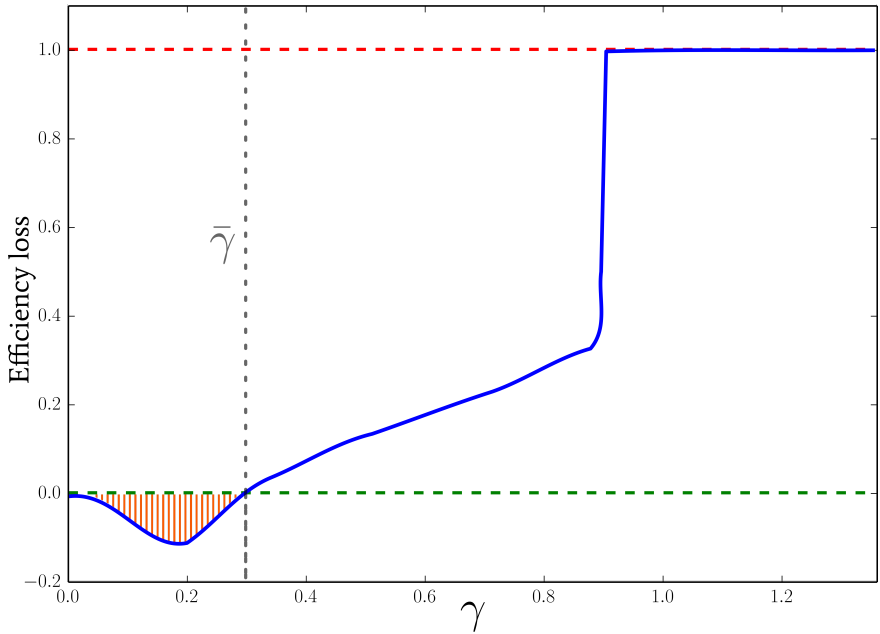


FIGURE 2.1: (Caption on the following page.)

FIGURE 2.1: Here we show the loss of efficiency in case of heterogeneous players with respect to the level of contribution that would have been achieved in the homogeneous case. A 100% loss (dashed red line) indicates that all players contribute nothing and thus that there is a complete loss of efficiency with respect to the homogeneous case. A 0% loss (dashed green line) means that the system reaches the same efficiency that it would have reached in the case of homogeneous players. A negative loss (values below the dashed green line) indicates that when endowments are heterogeneous the equilibrium reaches a higher efficiency than in the homogeneous case. As predicted, for superlinear payoffs $\gamma \geq 1$ the only possible equilibrium is non contribution by all and thus the loss of efficiency is total (right side of the picture). For intermediate sublinear payoff we can see that the efficiency is not completely lost and that it goes from being quite low to being closer to the homogeneous case. The quantitative value of efficiency that the mixed equilibrium achieves depends on the value of the mpcr and the width of the distribution of initial wealth as well as from other parameters. Finally, for $\gamma \leq \bar{\gamma}$ we first observe a slight increase in efficiency (below the dashed green line) and then the efficiency approaches the homogeneous one (left side of the picture), on account of the benefits of cooperation being obtainable for an arbitrary small contribution. Hence, for a wide range of values of γ , we observe a significant loss in efficiency compared to the homogeneous case. The simulation was obtained for the following set of parameters: $N = 100$, $S = 4$, $Q = 0.6$, $W_0 = 2$, $EN = 50$, $p = 0.2$ and $\sigma = 0.45$. For these parameters, $\bar{\gamma} \approx 0.18$.

For different values of the marginal per capita rate of return we observe (see figure A.2 in Appendix A) that the higher the mpcr, the wider is the area with partial efficiency losses in the picture and the smaller is the gain in efficiency around $\bar{\gamma}$.

Hence we can conclude that for a wide range of payoffs, the more realistic assumption of heterogeneous players leads to a disruptive loss in efficiency when compared to the homogeneous case. For a very limited range of γ , however, heterogeneity seems to result in a small increase in efficiency.

2.4 SUMMARY OF RESULTS

In summary, mechanisms based on assortative matching promise large efficiency gains when the interaction is such that it is safe to assume that the population consists only of equals. With heterogeneity, however, whether and how much assortative matching is likely to gain the population relative to random matching depends crucially on the provision efficacy of the public good and on the precise degree of heterogeneity. This implies that guarantees of more equal playing fields in these environments may be as important as implementation of assortative matching.

2.5 APPENDIX: NASH EQUILIBRIA

In this section we derive the Nash Equilibria of the generalized assortative matching voluntary contribution game.

We first define the game and derive some useful properties of it. Later, we show which equilibria exist for homogeneous and heterogeneous players for different values of the public good game efficacy.

NOTATION

- The expected payoff of player i is

$$E_i(\alpha_i, \alpha_{-i}) = (w_i - s_i) + \sum_{k=1}^M Pr(k | \alpha_i, \alpha_{-i}) \cdot \left[Q \cdot \left(S_{-i}^k + w_i \alpha_i^\gamma \right) \right] \quad (2.2)$$

with M the number of groups in the population, S_{-i}^k the sum of the effective contributions in agent's i group minus his own and $Pr(k | \alpha_i, \alpha_{-i})$ the probability of being ranked k th given α_i and α_{-i} . We indicate with \mathcal{S} the size of the groups.

- We say that players i, j are in class C_r if $s_i = s_j = s^r$. We write $c_r \equiv |C_r| = D_r \cdot \mathcal{S} + \tilde{c}_r$; $D_r, \tilde{c}_r \in N \cup \{0\}$.⁸ Note that by definition $0 \leq \tilde{c}_r < \mathcal{S}$.
- We call h the highest effective contribution and H the number of players s.t. $\alpha_i = h$; hence $H \equiv |C_1|$.
- We indicate with z the number of players playing the strategy $\bar{\alpha}$.
- Full heterogeneity means that $w_i \neq w_j \quad \forall i \neq j$.

Let us first compute the within group best response.

Lemma 2.5.1. *The best response within a group is: $\alpha_i = \bar{\alpha}$ if $0 < \gamma < 1$ and $\alpha_i = 0$ if $\gamma \geq 1$.*

Proof. The within group payoff is defined as following:

$$\phi_i(\alpha_i | \alpha_{-i}) = w_i(1 - \alpha_i) + Q \sum_{j \in G_i} \alpha_j^\gamma w_j$$

with G_i being the group to which agent i belongs and $\alpha_i \in [0, 1]$.

⁸ Or alternatively, D_r and \tilde{c}_r are defined as the unique non-negative integers such that $|C_r| = D_r \cdot \mathcal{S} + \tilde{c}_r$

The first order condition reads:

$$\frac{\partial \phi_i}{\partial \alpha_i} = -w_i + Qw_i\gamma\alpha_j^{\gamma-1} \stackrel{\text{FOC}}{=} 0 \Rightarrow \alpha_i = \left(\frac{1}{Q\gamma}\right)^{\frac{1}{\gamma-1}} \equiv \bar{\alpha} \quad (2.3)$$

implying the following payoff for agent i :

$$w_i(1 - \bar{\alpha}) + w_i Q \bar{\alpha}^\gamma + C \quad (2.4)$$

with C defined as $\sum_{j \in G_i, j \neq i} \alpha_j^\gamma w_j$.

The payoff for the corner strategies instead is $w_i + C$ for $\alpha = 0$ and $w_i Q + C$ for $\alpha = 1$.

Since $0 < Q < 1$ the payoff for $\alpha = 0$ is always bigger than the one $\alpha = 1$. Now we need to check when $\phi_i(\bar{\alpha}) > \phi_i(0)$ and hence when

$$Q \left(\frac{1}{Q\gamma}\right)^{\frac{\gamma}{\gamma-1}} > \left(\frac{1}{Q\gamma}\right)^{\frac{1}{\gamma-1}}.$$

If $0 < \gamma < 1$ we can rewrite the above expression as $Q(Q\gamma)^{\frac{\gamma}{1-\gamma}} > (Q\gamma)^{\frac{1}{1-\gamma}}$ and hence as $\gamma^\gamma > \gamma$; a condition that is always true for $0 < \gamma < 1$. If $\gamma > 1$ we instead obtain the condition $\gamma^{-\gamma} > \gamma^{-1}$, never true for $\gamma > 1$.

If $\gamma = 0$, the FOC trivially results in $\alpha = 0$.

Hence, the best response for player i if the group placement is independent from α_i is $\alpha_i = \bar{\alpha}$ for $0 < \gamma < 1$ and $\alpha_i = 0$ for $\gamma > 1$. \square

NE for homogeneous players

Here we compute what are the Nash Equilibria in the case of homogeneous players. The proof presented here follows closely the one in [38], changing only to adapt it to the generalized payoff. Note that for homogeneous players we have $w_i = w \forall i$.

We first of all note that, for homogeneous players, all players playing the within group Nash Equilibria is a best response.

Lemma 2.5.2. $\alpha_i = \bar{\alpha} \forall i$ is a Nash Equilibrium for every value of γ .

Proof. This is obviously an equilibrium. Since the mpcr Q is smaller than 1, there is no profitable deviation in being the only one contributing more than $\bar{\alpha}$. A player i deviating to a higher contribution would have the guarantee to be placed in the best group. The best group, however, would be such only because of him, thus making it not profitable to deviate. \square

In order to prove the existence of the high contribution equilibrium, we first observe that in order for the equilibrium to exist, the following conditions have to hold:

Lemma 2.5.3. *If there are some strategies α_i s.t. $\alpha_i > \bar{\alpha}$, then in equilibrium we have to have $H > S$ and $(H \bmod S) > 0$. Hence a player contributing more than $\bar{\alpha}$ has a non-zero probability to be grouped with a player contributing less than him.*

Proof. If the number of the highest contributors were a multiple of the group size, one of those player could unilaterally deviate and reduce his contribution by an amount small enough to remain in the same group and still profit from the deviation. For the same reason, that number of players has to be bigger than S . If it were smaller, in fact, high contributors would benefit by deviating to $\bar{\alpha}$. \square

Lemma 2.5.4. *If there are some strategies α_i s.t. $\alpha_i > \bar{\alpha}$, then the highest contribution h cannot be smaller than w . I.e. $\alpha_i = 1 \forall i \in C_1$.*

Proof. We call α_h the strategy s.t. $\alpha_h w = h$. From lemma 2.5.3 we know that (one of) the highest contributor(s) i has a non-zero probability to be grouped with some other player contributing less. Hence, if agent i were playing $\alpha_h < 1$, he could increase his contribution by an arbitrary small amount and be placed for certain in the best group.

Indeed the expected payoff for player i playing α_h is at most:

$$E_i(\alpha_h) < w(1 - \alpha_h) + QS w \alpha_h^\gamma \left(1 - \frac{\tilde{c}_1}{c_1}\right) + Qw [\alpha_h^\gamma \tilde{c}_1 + l^\gamma (S - \tilde{c}_1)] \quad (2.5)$$

where l is the strategy played an agent $j \in C_2$ and hence $l < \alpha_h$. \square

By deviating, player i would surely be placed in the best group, gaining

$$E_i(\alpha_h + \varepsilon) = w(1 - \alpha_h - \varepsilon) + QS w \alpha_h^\gamma + O(\varepsilon) \quad (2.6)$$

We have that $E_i(\alpha_h) < E_i(\alpha_h + \varepsilon)$ if

$$\varepsilon < Q \frac{\tilde{c}_1}{c_1} [(S - \tilde{c}_1) (\alpha_h^\gamma - l^\gamma)]$$

that for a small enough ε is true. Hence player i would be better off deviating to $\alpha_h + \varepsilon$. Consequently, $\alpha_i = 1 \forall i \in C_1$.

Lemma 2.5.5. *If there are some strategies α_i s.t. $\alpha_i > \bar{\alpha}$, then there cannot be any player j contributing $\bar{\alpha} < \alpha_j < 1$.*

Proof. Let us call j the player with the highest contribution after all the players contributing h ; i.e. $j \in C_2$ and let us call his strategy α_j .

If there are no ties regarding group membership, j could reduce his contribution to $\alpha_j - \varepsilon$ and remain in the same group.

If there are ties, j could increase his contribution by an arbitrary small amount and be sure to be grouped together with players belonging to C_1 . Similarly to lemma 2.5.4, it is possible to prove that for an arbitrary small ε we have:

$$E(\alpha_j) < E(\alpha_j + \varepsilon)$$

and hence that α_j cannot be an equilibrium strategy. \square

Lemma 2.5.6. *If there are some strategies α_i s.t. $\alpha_i > \bar{\alpha}$, then the number of players playing the within group best response is smaller than the group size; i.e. $z < S$.*

Proof. From lemma 2.5.3 we know that $(H \text{ mod } S) > 0$ and hence $((N - H) \text{ mod } S) > 0$.

If z were bigger than the group size, a player i contributing $\bar{\alpha}$ could increase his payoff by an arbitrary small amount and be placed with certainty in the group containing some players contributing h .

Hence there is a profitable deviation and $z \geq S$ could not be a Nash Equilibrium. \square

From lemmata 2.5.3-2.5.6 follows that if an equilibrium s.t. $\alpha_i > \bar{\alpha}$ for some i exists, then each player plays either $\bar{\alpha}$ or 1. Furthermore, the number of players playing the within group best response is smaller than the group size.

From the lemmata above we can derive under which condition the generalized voluntary contribution game has a Nash Equilibrium with high contributions level. Hence the existence of a nearly-efficient high equilibrium depends on the marginal per capita rate of return, the number of players and the size of the groups.

Theorem 2.5.7. *For values of γ bigger equal than a threshold value $\bar{\gamma}(Q, N, S)$, the generalized voluntary contribution game has Nash Equilibria in which $z < S$ players contribute $\bar{\alpha}$ and all the others $N - z$ players contribute 1. These equilibria are in addition to the equilibrium where all players contribute $\bar{\alpha}$.*

For $\gamma < \bar{\gamma}$, the only NE is that all players contribute the within group best response $\bar{\alpha}$.

Proof. For $N - z$ players contributing 1 to be a NE, we have to show that no full contributor has a profitable deviation to contribute $\bar{\alpha}$ and that no player contributing $\bar{\alpha}$ has an incentive to play 1. We write

$$E_1(1) = w \frac{\mathcal{S} - z}{N - z} Q [(\mathcal{S} - z) + z\bar{\alpha}^\gamma] + \frac{N - \mathcal{S}}{N - z} QSw \quad (2.7)$$

$$E_1(\bar{\alpha}) = w(1 - \bar{\alpha}) + wQ[\mathcal{S} - z - 1 + (z + 1)\bar{\alpha}^\gamma] \quad (2.8)$$

$$E_{\bar{\alpha}}(\bar{\alpha}) = w(1 - \bar{\alpha}) + wQ[\mathcal{S} - z + z\bar{\alpha}^\gamma] \quad (2.9)$$

$$E_{\bar{\alpha}}(1) = w \frac{\mathcal{S} - z + 1}{N - z + 1} Q [(\mathcal{S} - z + 1) + (z - 1)\bar{\alpha}^\gamma] + \frac{N - \mathcal{S}}{N - z + 1} QSw \quad (2.10)$$

where we indicate with $E_1(\alpha)$ the payoff of a high contributor and with $E_{\bar{\alpha}}(\alpha)$ the payoff of a low contributor.

For the above to be a NE we have to have that (2.7) > (2.8) and (2.9) > (2.10).

The first condition is equivalent to:

$$z \geq \frac{(1 - \bar{\alpha})N - QN(1 - \bar{\alpha}^\gamma)}{Q(1 - \bar{\alpha}^\gamma)[N - 1 - \mathcal{S}] + 1 - \bar{\alpha}} \quad (2.11)$$

The second condition leads to:

$$z \leq 1 + \frac{(1 - \bar{\alpha})N - QN(1 - \bar{\alpha}^\gamma)}{Q(1 - \bar{\alpha}^\gamma)[N - 1 - \mathcal{S}] + 1 - \bar{\alpha}} \quad (2.12)$$

However, it is important to remember that z needs to be smaller than the size of the groups.

Hence, we have that for values of the exponent γ such that (2.11) is at most $\mathcal{S} - 1$, the generalized voluntary contribution game has nearly efficient Nash Equilibria. For values of γ s.t. (2.11) is bigger than $\mathcal{S} - 1$ the only equilibrium is the one where all players play the within group best response. We call $\bar{\gamma}$ the value of γ such that eq. (2.11) = $\mathcal{S} - 1$. □

Hence we obtain the existence of a nearly-efficient high equilibrium depends on the marginal per capita rate of return, the number of players and the size of the groups.

NE for heterogeneous players

Here we show that the near efficient Nash Equilibrium cannot exist for heterogeneous players. Furthermore, we derive under which condition the generalized voluntary contribution game has a pure strategy NE.

In the following we prove the lemmata necessary to derive the equilibrium.

Lemma 2.5.8. *In case of full heterogeneity ⁹ the assumptions of Lemma 2.5.3 cannot be satisfied.*

Proof. Let's assume that they are and show that this cannot be a NE. We call k the player with the lowest w_i belonging to C_1 .

We have two possibilities: (a) $s_k = w_k$ and (b) $s_k < w_k$

- (a):

Let's take a player j s.t. $j \in C_1$ and $j \neq k$. When playing $s_j = s^1$, he has an expected payoff of at most (because in the mixed group that could be also player of classes lower than 2):

$$E(j) \leq w_j - s^1 + ms^1 \mathcal{S} \cdot \left(1 - \frac{\tilde{c}_1}{c_1}\right) + m \left[s^1 \tilde{c}_1 + s^2 (\mathcal{S} - \tilde{c}_1)\right] \cdot \frac{\tilde{c}_1}{c_1} \quad (2.13)$$

where $\frac{\tilde{c}_1}{c_1}$ is the probability that agent j ends up in the group where not all agents belong to C_1 .

But agent j could play α_ε s.t. $s_j = s^1 + \varepsilon$, being guaranteed to end up in the first group and thus having an expected payoff of

$$E_\varepsilon(j) = w_j - s^1 - \varepsilon + m (\mathcal{S} - 1) s^1 + m (s^1 + \varepsilon)$$

But if $\varepsilon < ms^1 (\mathcal{S} - \tilde{c}_1) \frac{\tilde{c}_1}{c_1} \left(1 - \frac{s_2}{s_1}\right)$ we have that $E_\varepsilon(j) > E(j)$. ¹⁰ Hence there is a profitable deviation for agent j ; so there can't be a NE.

⁹ Actually it is already valid for "enough" heterogeneity.

¹⁰ Because:

$$\begin{aligned} A &= m \left[s^1 \mathcal{S} + \varepsilon\right] - \varepsilon > ms^1 \mathcal{S} \cdot \left(1 - \frac{\tilde{c}_1}{c_1}\right) + m \left[s^1 \tilde{c}_1 + s^2 (\mathcal{S} - \tilde{c}_1)\right] \cdot \frac{\tilde{c}_1}{c_1} \\ A &> ms^1 \mathcal{S} + \varepsilon > ms^1 \mathcal{S} \cdot \left(1 - \frac{\tilde{c}_1}{c_1}\right) + ms^1 \mathcal{S} \cdot \frac{\tilde{c}_1}{c_1} + m \left[s^1 (\tilde{c}_1 - \mathcal{S}) + s^2 (\mathcal{S} - \tilde{c}_1)\right] \cdot \frac{\tilde{c}_1}{c_1} \\ -\varepsilon &> ms^1 (\tilde{c}_1 - \mathcal{S}) \frac{\tilde{c}_1}{c_1} \left[\frac{s^2}{s^1} - 1\right] \end{aligned}$$

- (b):

The same as (a), except that now even for k it is profitable to deviate to $s_k = s^1 + \varepsilon$.

Hence for heterogeneous players, it is impossible to maintain the conditions under which nearly efficient equilibrium was possible. \square

Lemma 2.5.9. *In case of full heterogeneity of endowments w_i , with all w_i having the same order of magnitude, $\alpha_i = \bar{\alpha} \forall i$ is not a Nash Equilibrium.*

Proof. If all players were playing $\bar{\alpha}$, they would be ranked based on their endowments. Hence, a player i with the biggest endowment w_i smaller than the biggest S endowments would have a profitable deviation by playing $\alpha_i = \bar{\alpha} + \varepsilon$ and be assigned to the best group.

If the endowments are such that $w_i < w_{i+1} \alpha_{i+1} \forall i$ ¹¹, then there are no profitable deviations and $\alpha_i = \bar{\alpha} \forall i$ is a Nash Equilibrium. \square

From the above lemmas we can derive the following theorem:

Theorem 2.5.10. *For $\gamma \geq 1$, the generalized voluntary contribution game has only equilibrium non-contribution by all. For $0 < \gamma < 1$, the game has no pure strategy NE and hence it has a mixed strategy Nash Equilibrium.*

Proof. From lemma 2.5.8 we know that the nearly efficient NE cannot exist for heterogeneous players. Furthermore, we can prove as in lemma 2.5.5 that there can be no pure strategies such that $\bar{\alpha} < \alpha_i < 1$ for any player i .

From lemma 2.5.1 we know that for $\gamma \geq 1$ the within group best response is to play $\alpha_i = 0 \forall i$. This can be an equilibrium and hence for values of γ bigger than 1 there is a unique pure strategy NE for the generalized voluntary contribution game.

For $\gamma < 1$, however, the within group best response is to play $\alpha_i = \bar{\alpha}$. But lemma 2.5.9 shows that this cannot be an equilibrium of the game (if the values of the endowment don't differ too much). Hence for $\gamma < 1$ there are no pure strategy NE and thus there has to exist a mixed strategy Nash Equilibrium. \square

¹¹ With the endowments ranked from the lowest to the highest.

CONTRIBUTION-BASED GROUPING UNDER NOISE

ABSTRACT

Many real-world mechanisms are “noisy” or “fuzzy”, that is the institutions in place to implement them operate with non-negligible degrees of imprecision and error. This observation raises the more general question of whether mechanisms that work in theory are also robust to more realistic assumptions such as noise. In this paper, in the context of voluntary contribution games, we focus on a mechanism known as “contribution-based competitive grouping”. First, we analyze how the mechanism works under noise and what happens when other assumptions such as population homogeneity are relaxed. Second, we investigate the welfare properties of the mechanism, interpreting noise as a policy instrument, and we use logit dynamic simulations to formulate mechanism design recommendations.

3.1 MOTIVATION

Typically, individual decisions in social-dilemma interactions are not perfectly observable in the real world. Applied mechanism designers should keep this in mind when implementing mechanisms, in particular in the context of interactions that have the strategic nature of voluntary contribution games where imperfect observability is ubiquitous.¹ Real-world institutions are usually “noisy” or “fuzzy”, operating with non-negligible degree of imprecision. By contrast, theory investigations of mechanisms for the most part study perfect mechanisms. In this chapter, as a first step towards performing robustness checks of mechanisms under relaxed assumptions more generally, we investigate the mechanism of contribution-based competitive grouping (as introduced by [38]): we relax some assumptions and investigate what happens when there is (i) noise, (ii) heterogeneity and (iii) different action spaces.

The chapter is structured as follows. Next, we introduce our model. In Section 3.3, we first analyze how the existence of Nash equilibria depends on (iii) the strategy space of the game and on (ii) the underlying population homogeneity/heterogeneity in terms of contribution budgets. Then, we assess their robustness when we allow more and more (i) monitoring noise. In Section 3.4, we investigate how a social planner, interpreting monitoring noise and/or the other model characteristics as policy instruments, would trade-off efficiency and equality to maximize social welfare. Finally, we use agent-based simulations to study the logit dynamics of the game in order to quantify the effect of noisy monitoring.

Our results are summarized as follows. In terms of the existence of Nash equilibria, the zero-contribution outcome is always a Nash equilibrium and becomes the unique one under too much noise. High-contribution equilibria exist if three things come together: the rate of return of the underlying contribution game is high enough, grouping is sufficiently precise, and agent homogeneity is established (either in terms of budgets or via the strategy space). In terms of welfare, implementations with an intermediate level of noise/fuzziness, enabling high efficiency gains at low inequality costs, maximize welfare in most cases. The exception is the case of heterogeneous budgets and binary (all-or-nothing) contribution decisions, where less noise is unambiguously preferable. Finally, our logit-dynamics simulations indicate that high-contribution equilibria, when existent, tend to be

¹ For other social-dilemma contexts, see, for example, [99, 100] for imperfect public monitoring, [101] for noisy prisoners’ dilemmas and [102] for team-production games with group-level information.

more robust than the zero-contribution equilibrium, and we are able to identify an optimal level of mechanism noise/fuzziness as a function of the behavioral noise that is inherent under the logit-choice rule.

3.2 THE MECHANISM

3.2.1 The Model

Population $N = \{1, 2, \dots, n\}$ plays the following three-step game under common knowledge.

Step 1. Voluntary contributions:

Action Space 1. Binary: Each $i \in N$ simultaneously decides whether to set $c'_i = 0$ (i.e., to free-ride) or to set $c'_i = 1$ (i.e., to contribute).

Action Space 2. Continuous: Each $i \in N$ simultaneously decides on a proportional contribution $c'_i \in [0, 1]$ between free-ride ($c'_i = 0$) and contribute ($c'_i = 1$).

Each $i \in N$ has a budget of $B_i \in \mathcal{R}^+$, and his/her above action results in an effective contribution of $c_i = c'_i \cdot B_i$. The effective contributions yield a population contribution vector $\mathbf{c} = \{c_i\}_{i \in N}$. Denote by \mathbf{c}_{-i} the vector excluding i .

Budget 1. Homogeneity: $B_i = 1$ for all i .

Budget 2. Heterogeneity: $B_i \neq B_j \forall i \neq j$ with (w.l.o.g.) $B_i < B_{i-1}$. We place a very mild constraint on how different endowments can be by imposing that $\exists X \in \mathcal{R}$ s.t. $B_i > B_{i-1} + X \forall i$ and any $X > 0$.

Step 2. Fuzzy grouping based on a ‘noisy’ ranking of contributions:

Step 2.1. From ‘noisy’ rankings...: Instead of c_i , the authority observes $x_i = c_i + e_i$, where e_i is some i.i.d. white noise with mean zero and variance $\sigma^2 = \frac{1-\beta}{\beta}$ where $\beta \in (0, 1]$. Based on the vector of observed contributions \mathbf{x} , players are ranked from highest to lowest.

β takes the role of a “meritocracy” parameter in our setting: (i) under no meritocracy (when $\beta \rightarrow 0$), all rankings are equally likely, and all players have the same expected rank²; (ii) under full meritocracy (when $\beta = 1$), only “perfect” rankings are possible, that is, players who contributed more have a higher rank than players who contributed less, and ties are ran-

² With a slight abuse of notation, from now on, we will write $\beta \in [0, 1]$ and $\beta = 0$ to indicate the “no meritocracy” case where $\sigma \rightarrow \infty$, and the ranking, and consequentially, the matching, of the players is (uniformly) randomly selected.

domly broken; (iii) under intermediate meritocracy (when $\beta \in (0, 1)$), all rankings have positive probability, but on aggregate, depending on β , players who contributed more have a higher expected rank than players who contributed less. As β increases, we move continuously from (i) to (ii).

Step 2.2. ...to fuzzy groupings: Groups form based on the noisy ranking such that g groups $\{S_1, S_2, \dots, S_g\}$ of a fixed size $s < n$ form. The result is a partition ρ of N (where $s = n/g > 1$ for some $s, g \in \mathbf{N}^+$) with groups $S_p \in \rho$ (s.t. $p = 1, 2, \dots, g$) where the s highest-ranked players form Group 1, the s second-highest players form Group 2, etc.³

(See Appendix 3.9 for implementation examples of fuzzy grouping.)

Step 3. Payoffs:

Step 3.1. Realized payoffs (ex post): Given \mathbf{c} and ρ , total payoffs generated in each $S \in \rho$ are $s + (r - 1) \sum_{i \in S} c_i$, where r is the rate of return. Each $i \in N$ s.t. $i \in S \in \rho$ receives an individual payoff of:

$$\phi_i(c_i | c_{-i}, \rho) = \underbrace{(1 - c_i)}_{\text{remainder from budget}} + \underbrace{(R) * \sum_{j \in S} c_j}_{\text{return from the public good}}, \quad (3.1)$$

where $R := \frac{r}{s}$ is the marginal per capita rate of return; it is standard to assume a ‘social dilemma’ character of the game by setting $R \in (\frac{1}{s}, 1)$.⁴ Let $\phi = \{\phi_i\}_{i \in N}$ denote the payoff vector.

Step 3.2. Expected payoffs (ex ante): Groups form in Step 2, and payoffs materialize in Step 3, both based on the sunk contribution decisions taken during Step 1. From expression (3.1), i ’s expected payoff of contributing c_i , as evaluated during Step 1 given c_{-i} , is therefore:

$$\begin{aligned} \underbrace{\mathbf{E}[\phi_i(c_i | c_{-i})]}_{\text{exp. return from } c_i} &= \underbrace{1}_{(i) \text{ budget}} - \underbrace{(1 - R) * c_i}_{(ii) \text{ sure loss on own contribution}} + \underbrace{R * \mathbf{E}\left[\left(\sum_{j \neq i: j \in S} c_j\right) \mid c_i\right]}_{(iii) \text{ exp. return from others in group}}, \quad (3.2) \end{aligned}$$

³ Meritocracy β guides smoothly from (i) no meritocracy (grouping is random as in [21]) to (ii) full meritocracy (the case of perfect contribution-based grouping as in [38]). Note that many public goods experiments use variants of Andreoni’s random (re-)matching implementation (e.g., [21, 42, 43, 103–110]); see [20, 93] for reviews.

⁴ Thus, full-contribution is collectively efficient, and zero-contribution, despite collectively inefficient, is the unique Nash equilibrium under no meritocracy.

Note that Term (iii), the expected return from others' contributions, may depend on one's own contribution if $\beta > 0$.

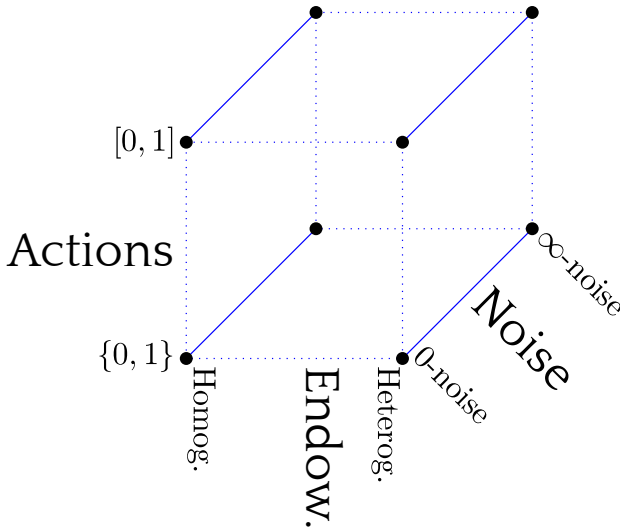


FIGURE 3.1: The figure illustrates the three dimensions that we vary to assess the robustness of the mechanism. Two are binary (indicated by the blue dotted lines): For one, players have either the same or different initial endowments; further, they either choose a proportional or an all-or-nothing contribution. One dimension is varied continuously (indicated by the blue continuous line): contributions are observed with different degrees of noise.

3.3 NASH EQUILIBRIA

Next, we analyze the Nash equilibria in terms of the ex ante decisions made in the games that we detailed in the previous section, where various games result from the different combinations of the chosen model ingredients. Figure 3.2 summarizes what is presented in more detail below.

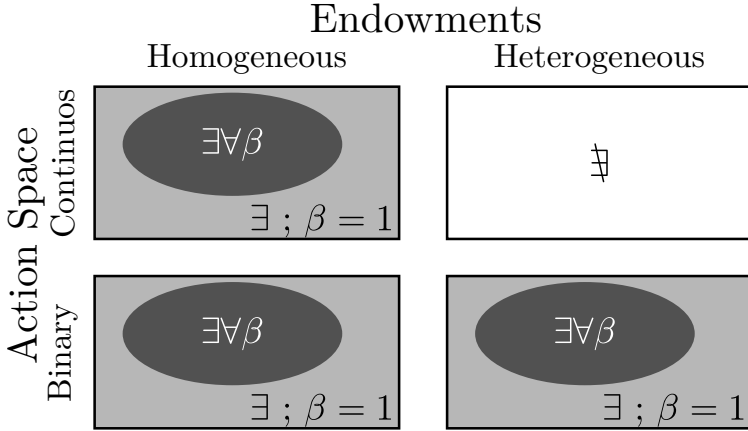


FIGURE 3.2: The figure summarizes the existence of the “high-efficiency” equilibria described in this section. For homogeneous endowments, regardless of the action space, there exist high contribution equilibria for high enough marginal rates of return. In the case of heterogeneous endowments, only a binary action space allows for contributive equilibria. Whenever high contribution equilibria exist under full meritocracy, there exists a marginal rate of return (higher than the one for $\beta = 1$) such that the high equilibria exist also with a fuzzy mechanism.

3.3.1 Game 1: ‘Baseline’

The following equilibrium structure results in the no-noise implementation with agent homogeneity and continuous action space (as in [38]). For this case, there always exists a Nash equilibrium with zero contributions; as in standard voluntary contribution games with a single group or with random grouping. We call this equilibrium the “non-contribution equilibrium”. However, non-contribution is not a dominant strategy, as it would be if groups were randomly assigned, because players may prefer to contribute when this gets them into high-contribution groups that promise higher payoffs. As a result, provided the marginal per capita rate of return R is high enough, [38] prove that there exist new asymmetric pure-strategy Nash equilibria where most players contribute and very few players free-ride.⁵ We call these equilibria “high-contribution equilibria”. We call the

⁵ See [38], Theorem 1.

threshold for which these equilibria exist *mPCR*; note that this threshold is increasing in the relative group size s/n , which implies that it is, *ceteris paribus*, harder (easier) to satisfy for relatively large groups (populations)

High-contribution equilibria are particularly interesting, because several recent experimental studies provide support for them [38, 111, 112].⁶

3.3.2 Game 2: 'Heterogeneity Extension' (Extension 1)

3.3.2.1 Heterogeneity and Continuous Action Space

The first negative result is that under heterogeneity, non-contribution is the only existing equilibrium. A key property of high-contribution equilibria in the baseline case is that there exists a mixed group where both defectors and contributors are placed. In these equilibria, a contributor has a high-enough probability to be placed in a group filled only with contributors and a low enough probability to be placed in the mixed group so that it is, in expectation, not beneficial to free-ride and be placed with certainty in the low group instead.

However, in chapter 2, it is proven that, under heterogeneity, there cannot exist an equilibrium where two or more players contribute the same amount. The reason is that, if two or more players contribute the same amount, then the one with the highest endowment would have a profitable deviation in contributing slightly more due to the existence of the mixed group: contributing only slightly more, and he/she is guaranteed not to be placed in a worse group. Conversely, if players in a higher group already contribute everything, then the wealthiest player could instead contribute slightly less and still be guaranteed to remain in the same group. Therefore, when players have access to a continuous action space and can thus reduce or increase their contributions by an infinitesimal amount, then there can be no high contribution equilibrium for heterogeneous endowments.

It is important to note that this result relies on players all having different endowments. Indeed, if players were to belong only to two endowment classes, Gunthorsdottir et al. [39] prove that there might exist an equilibrium where every player contributes everything. This equilibrium, however, requires conditions that become harder to satisfy the more classes of endowments are introduced, eventually becoming impossible if there are as many levels of endowment as there are players. Interestingly, Gunthors-

⁶ In fact, recent evidence suggests that players may endogenously implement variants of contribution-based competitive grouping over time and then converge to the high-contribution equilibria [113].

dottir et al. [39, 111] also find empirical support for the high-efficiency equilibria by showing that, when existing, near-efficient equilibria are preferred over the worst ones.

3.3.2.2 *Binary Action Space*

With a binary action space, the above argument about infinitesimal changes in contributions does not hold any more. Indeed, it can be shown (see Proposition 3.7.5 in Appendix 3.7) that in the case of heterogeneous endowments and binary action space, there exists a threshold for the marginal rate of return such that contributive equilibria exist.

For this configuration, a maximum of g pure strategy Nash equilibria can coexist, depending on the value of r . Their structure is so that the $k \cdot s$ ($0 \leq k < g$) poorest players defect and the $(g - k) \cdot s$ richest players contribute. The non-contribution equilibrium ($k = 0$) always exists, regardless of the value of the marginal rate of return. Interestingly, for the equilibrium with $k = 1$ to exist, the marginal rate of return has to be higher than for the equilibrium with $k = 2$ to exist, and so on, until the lowest threshold for R such that the only contributive equilibrium is the one where all but one group is filled with contributors⁷.

As a corollary of Proposition 3.7.5, there also exist high-contribution equilibria for homogeneous endowments and binary action space. In this case, all thresholds collapse and reduce to the same value as in the baseline case, and all the existing equilibria are such that there exists a mixed group of contributors and defectors, in analogy with the continuous action space case.

3.3.3 *Game 3: 'Noise Extension' (Extension 2)*

In the case of a noisy mechanism, the existence of Nash equilibria summarizes as follows.

Whenever the chosen model allows for high-contribution equilibria in perfect meritocracy, then there exists a noise/fuzziness level below which the same equilibria continue to exist: given any $R > mpcr$, there exists a $\beta < 1$ such that the same high-contribution equilibria (possibly with more free-riders than under $\beta = 1$, but $< s$) continue to exist as when $\beta = 1$ (see Appendix 3.7, Proposition 3.7.8). The minimum level of β , denoted by $\underline{\beta}$,

⁷ For a marginal per capita rate of return $R > s$, the game is not a social dilemma anymore: "cooperate" becomes a dominant strategy, and the only existing equilibrium is for everybody to contribute everything.

for which high-contribution equilibria exist, is an implicit function that is decreasing in R and increasing s/n .

Universal non-contribution is a Nash equilibrium for any β (see Appendix 3.7, Proposition 3.7.7). The reason for this is that the only incentive for an individual to contribute a positive amount under contribution-based grouping is to be grouped with others doing likewise. When no one else contributes, there is no such incentive.

3.3.4 Remark: Mixed-Strategy Nash Equilibria

With heterogeneous endowments, symmetric mixed-strategy NE do not exist, either because the only existing equilibrium is non-contribution by all or because of the structure of the high-efficient equilibria (see Proposition 3.7.5 in Appendix 3.7 and Theorem 2 in chapter 2).

Similarly, for heterogeneous endowments and continuous action space, mixed-strategy equilibria do not exist (see Appendix 3.8, Corollary 3.8.3.2). When contribution decisions are restricted to coarse or binary action spaces, however, it is the case that, for every β , there exists a $R \in (mocr, 1)$ at which there exists one and above which there exist two mixed-strategy Nash equilibria of the form ‘contribute fully with p and free-ride with $1 - p$ ’, one with a high contribution probability \bar{p} and one with a low p (see Appendix 3.8, Proposition 3.8.1).⁸

3.4 WELFARE COMPARISON

We now turn to the point of view of a social planner and analyze the welfare properties of equilibria in terms of realized payoffs ϕ .⁹ Ex post, in contrast to ex ante, it is not always the case that high-contribution equilibria are payoff-dominant.¹⁰ Hence, it is not obvious which equilibria a social planner who cares about equality will prefer. To evaluate welfare, we therefore use a variant of the social welfare function of [117], which has the advantage that, like the parameter β governing meritocracy, a single

⁸ Note that, whenever pure strategy highly-efficient equilibria exist, there could also exist several asymmetric Nash equilibria whose characterization is not easily obtained.

⁹ Note that Harsanyi’s social welfare approach [114], by contrast, would consider ex ante payoffs, that is expected payoffs. His social welfare function is $W_H(\phi) = \frac{1}{n} \sum_{i \in N} \mathbf{E}[\phi_i]$. See, for example, [115] for a discussion of ex ante versus ex post approaches.

¹⁰ The work in [116] defines that outcome ϕ payoff-dominates ϕ' if $\phi_i \geq \phi'_i$ for all i , and there exists a j such that $\phi_j > \phi'_j$.

parameter characterizes a continuous range of social planner preferences, spanning preferences across the entire equality-efficiency tradeoff.

Social welfare: Given outcome (ρ, ϕ) , let $W_e(\phi)$ be the social welfare function measuring its welfare given the inequality aversion parameter $e \in [0, \infty)$:

$$W_e(\phi) = \frac{1}{n(1-e)} \sum_{i \in N} \phi_i^{1-e} \quad (3.3)$$

(For $e = 1$, expression (3.3) is defined by $W_1(\phi) = \frac{1}{n} \prod_{i \in N} \phi_i$, i.e., by the Nash product.) Expression (3.3) nests both the utilitarian (Bentham) and Rawlsian social welfare functions as special cases.¹¹ When $e = 0$, Expression (3.3) reduces to $W_0(\phi) = \frac{1}{n} \sum_{i \in N} \phi_i$, i.e., the utilitarian criterion measuring the state's 'efficiency' (total payoffs) only. When $e \rightarrow \infty$, Expression (3.3) approaches $W_\infty(\phi) = \min(\phi_i)$, i.e., the Rawlsian criterion measuring the outcome's welfare by the utility of the worst-off. Obviously, a utilitarian social planner prefers the high-contribution equilibria to the non-contribution equilibrium, because it is more efficient. A Rawlsian, however, would prefer the non-contribution equilibrium (with perfect equality of payoffs) if any player receives a lesser payoff in the high-contribution equilibria.

Which equilibrium is preferable in terms of social welfare for any given social welfare function depends on the game considered and on the social planner's relative weights on efficiency and equality. Critical for this assessment is the inequality aversion e .

Welfare criterion: The social planner acts in order to maximize $\mathbf{E}[W_e(\phi)]$, where ϕ are evaluated according to the equilibrium selection criterion. Suppose that the social planner can set some of the rules of the game to achieve the above goal: Which conditions maximize the social welfare? Depending on the game, what is the value of $\beta \in [0, 1]$ that maximizes $\mathbf{E}[W_e(\phi)]$?

3.4.1 *Homogeneous Endowments*

In case of homogeneous endowments, there is a clear trade-off between efficiency and equality: the high gains in efficiency are obtained at the expense of the cooperative players placed in the group containing some defectors. These few players are worse off than in the case where everybody defects. A very inequality averse social planner might choose to increase the fuzziness in the observations in order to revert to the non-contribution equilibrium and thus improve the ex-post payoff of the worst off.

¹¹ See, for example, [118] for a discussion of this generalization.

For an example, consider the economy illustrated in Table 3.1 (with $n = 16$, $s = 4$ and $r = 1.6$), and suppose a social planner considers moving from $\beta = 0$ to $\beta = 1$ (i.e., from completely random to fully-meritocratic ranking). At $\beta = 0$, he/she considers the non-contribution equilibrium. At $\beta = 1$, he/she considers the high-contribution equilibrium. To assess which one he/she prefers, he/she makes a W_e -comparison. For this numerical example, it turns out that for any W_e with $e < 10.3$ he/she prefers the high-contribution equilibria, while for a W_e with $e \geq 10.3$, he/she prefers the non-contribution equilibrium.¹² The following general statement about the welfare-optimal β can be made.

Proposition 3.4.1. *For any $R > \max\{mpcr, 1/(s-1)\}$, there exists a population size $n < \infty$ such that $\mathbf{E}[W_e(\phi); \underline{\beta}] > \mathbf{E}[W_e(\phi); \beta]$ for all $\beta \neq \underline{\beta}$ given any parameter of inequality aversion e bounded away from ∞ .*

Proof. Below $\underline{\beta}$, the high-contribution equilibrium does not exist. The non-contribution equilibrium cannot be welfare-maximizing for any $e < \infty$ as $n \rightarrow \infty$, because large efficiency gains (approaching infinity) would be foregone at virtually no inequality cost. Consider setting some $\bar{\beta} \in (\underline{\beta}, 1)$ instead, for which the high-contribution equilibrium exists. Write q_1^n for the probability of having more than one free-rider in any group for a realized outcome (ρ, ϕ) given $n < \infty$. Since the number of free-riders does not increase as n increases, $\partial q_1^n / \partial n < 0$. Since contributors in groups with at most one free-rider receive a payoff strictly greater than one ($(s-1)R > 1$), we have $\mathbf{E}[W_e(\phi); \bar{\beta}] > (1 - q_1^n) \times W_e(\phi_i = (s-1)R \forall i)$. Because, given any $\beta < 1$, $\partial q_1^n / \partial n < 0$, there therefore must exist a population size $n < \infty$ above which $\mathbf{E}[W_e(\phi)] > W_e(\phi_i = 1 \forall i)$. Hence, $\underline{\beta}$ is generally welfare-optimal for any $e < \infty$ as $n \rightarrow \infty$. \square

Remark 3.4.2. $\mathbf{E}[W_e(\phi); \underline{\beta}] > \mathbf{E}[W_e(\phi); \beta]$ for all $\beta \neq \underline{\beta}$ is also the case for n bounded away from infinity if (a) e is set below some bound $e < \infty$ and/or (b) R is set above some bound $R > 1/(s-1)$.

¹² With $e = 10.3$, W_e requires efficiency gains of more than twice the amount lost by any player to compensate for the additional inequality.

High-Contribution Equilibrium when $\beta = 1$	Payoff	Non-Contribution Equilibrium when $\beta = 0$
	0	0
	0	0
	0	0
	0	0
13 14 ($c_i = 1$)	2	0
	0	16 ($c_i = 0$)
	0	0
	0	0
	0	0
1 2 3 4 5 6 7 8 9 10 11 12 ($c_i = 1$)	12	0
15 16 ($c_i = 0$)	2	0
	24.4 efficiency	16

The stem of the table comprises the payoffs. The leafs are the number of players receiving that payoff (with their contribution decision) and the individual ranks of players corresponding to payoffs in the two equilibria. At the bottom, the efficiencies of the two outcomes are calculated.

TABLE 3.1: Stem-and-leaf plot of individual payoffs for the non-contribution equilibrium (valid for any $\beta \in [0, 1]$) and for the high-contribution equilibrium (when evaluated at $\beta = 1$). Parameter values are $n = 16$, $s = 4$, $r = 1.6$ and $\beta = 1$.

3.4.2 Heterogeneous Endowments

In the case of heterogeneous endowments and binary action space¹³, there is no equilibrium in which some contributors are grouped together with some defectors. Hence, no player is worse off in the high-contributing equilibria compared to the non-contributing one. Indeed, every contributive equilibria Pareto dominates the non-contributing one. The equilibrium where all but the S poorest players contribute is ex-post payoff dominant.

Hence, regardless of the value of the social planner preference parameter e , a social planner will always prefer the least possible amount of noise in the observations (ideally $\beta = 1$) in order to ensure the most efficient outcome.

Note however that this does not mean that inequality does not increase due to the agents playing the efficient equilibrium. The starting inequality due to the initial endowments is amplified by the fact that the poorest players are the ones in the non-contributing group and thus the ones

¹³ For continuous action space, the only equilibrium is non-contribution by all, and therefore, there is no social welfare analysis to be made.

that do not benefit from the common good. Indeed, if the social planner were to measure inequality based on statistical dispersion measures rather than looking at the worst off player, he/she would observe an increase in inequality. As an example, consider the Gini coefficient¹⁴, a standard measure of inequality in a population, for a population of 16 players with initial endowments distributed according to a normal distribution $\mathcal{N}(1, 0.1^2)$: for a realization of the starting endowment resulting in a Gini coefficient of ~ 0.09 , the Gini coefficient for the realized payoff in the ex-post dominant equilibrium is ~ 0.18 . Hence, by this measure, the inequality in the population doubled after the game.

3.5 LOGIT DYNAMICS

In order to quantify the effect of noise, we simulate the above games with an agent-based model where players use logit-response [81, 121] to study the dynamics of the game. The following algorithm is used in these simulations:

¹⁴ For a definition of the Gini coefficient, see, e.g., [119, 120]

Algorithm 1: Logit choice dynamics.

- 1 Initialization: Assign the initial endowments to each player sampling them from a Gaussian distribution with mean W_0 and standard deviation σ . If some endowments are zero or negative, the sampling is repeated.^a
 - 2 Initialization: Set initial strategy α_i to 0 $\forall i$ (start the simulation in the fully-defective state^b), and set time t to zero.
 - repeat**
 - 3 Update players' strategy:
 - for** $i = 1$ to N **do**
 - Generate random number X uniformly in $[0, 1)$.
 - if** $X < p$ **then**
 - Do not update player i 's strategy.^c
 - else**
 - Player i chooses a new strategy j based on the logit probability distribution^d: $P_j = \frac{\exp(\lambda EU_j(\alpha_{-i}))}{\sum_k \exp(\lambda EU_k(\alpha_{-i}))}$, where λ is the rationality parameter and $EU_j(\alpha_{-i})$ is the expected utility of strategy j given the α_{-i} , the strategies played by all the other players in the previous round.
 - 4 Matching: Based on the contribution of each player, groups are formed as described in Section 3.2.1 and payoffs are then materialized.
 - 5 Reset wealth: Set the wealth of each player to his/her initial one and increase time t by one.
 - until** $t < T$;
-

^a The standard deviation is chosen for this to be an unlikely event.

^b Different initial starting conditions have been explored, and they have been observed to have no effect on the final outcome of the simulation.

^c Inertia was added to ensure the convergence to the pure-strategy Nash equilibrium (if existing). Without inertia, the best-response dynamics could oscillate around such an equilibrium.

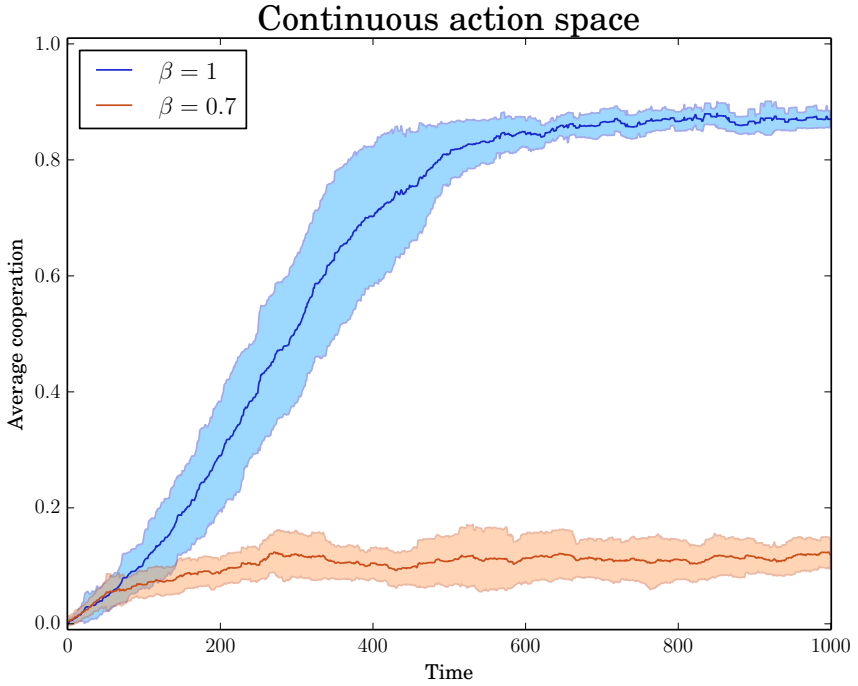
^d For a definition, see, e.g., [82].

After T rounds, the algorithm stops, and the population average of the strategies is computed. The procedure is repeated EN times, and the ensemble average of the population average is obtained.

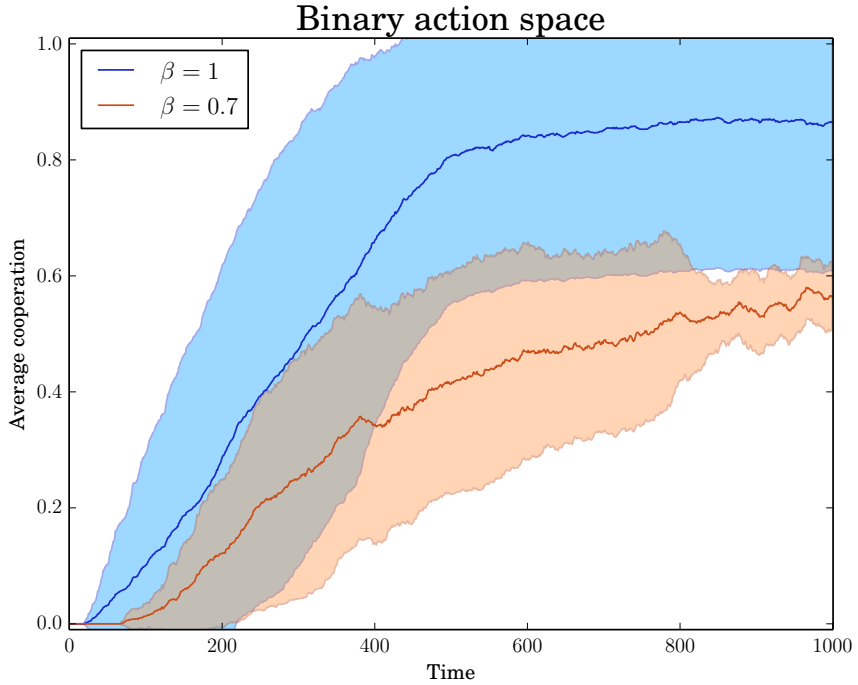
Figure 3.3 shows example results for homogeneous endowments in four different cases: perfect and imperfect observations for continuous and bi-

nary action spaces. When the matching mechanism is implemented perfectly, we observe a high level of contributions both in the binary and continuous action space (blue lines in Figure 3.3a,b). However, for an intermediate level of noise in the observed contributions, the results are quite different: in the binary case, we observe an intermediate level of average contributions (orange line in Figure 3.3b), while in the continuous case, we see very low amounts of cooperation (orange line in Figure 3.3a).

In Figure 3.4, we quantify how the percentage of contributions depends on the rate of return R and the level of noise in the matching mechanism β for agents with homogeneous endowments in a strategy space with continuous action. As predicted, the average level of cooperation weakly increases with increasing R and β , and we observe almost complete cooperation when approaching one. Naturally, regardless of the value of β , we observe no cooperation for a value of $R \leq \frac{1}{s}$. The contour lines indicate when the average contribution is above a certain threshold.



a Average cooperation in continuous action space. These simulations are obtained with $N = 80$ and $\lambda = 10$.



b Average cooperation in binary action space. These simulations are obtained with $N = 16$ and $\lambda = 20$.

FIGURE 3.3: (Caption on the following page.)

FIGURE 3.3: The figure summarizes the average level of contributions in continuous and binary action spaces. The average is shown as a function of time for the voluntary contribution game with meritocratic matching with homogeneous endowments in four different cases: perfect and imperfect observations for continuous (a) and binary (b) action spaces. When the matching mechanism is implemented without noise ($\beta = 1$, blue lines in (a,b)), we observe high levels of cooperation in both cases. On the other hand, when the matching mechanism is implemented with partially-noisy observations $\beta = 0.7$, orange lines in (a,b), the result depends on the strategy space of the game: for a continuous action space (a), we observe a very low percentage of average contributions, while for a binary action space (b), we see a higher level of cooperation. The blurred areas around the lines represent the 95% confidence interval. The simulations were obtained for the following set of parameters: $s = 4$, $R = 0.5$, $EN = 200$, $p = 0.99$. The simulations for the binary action space are obtained with a lower value for the rationality parameter in the logit function in order to speed up convergence and with a higher number of agents to reduce the error.

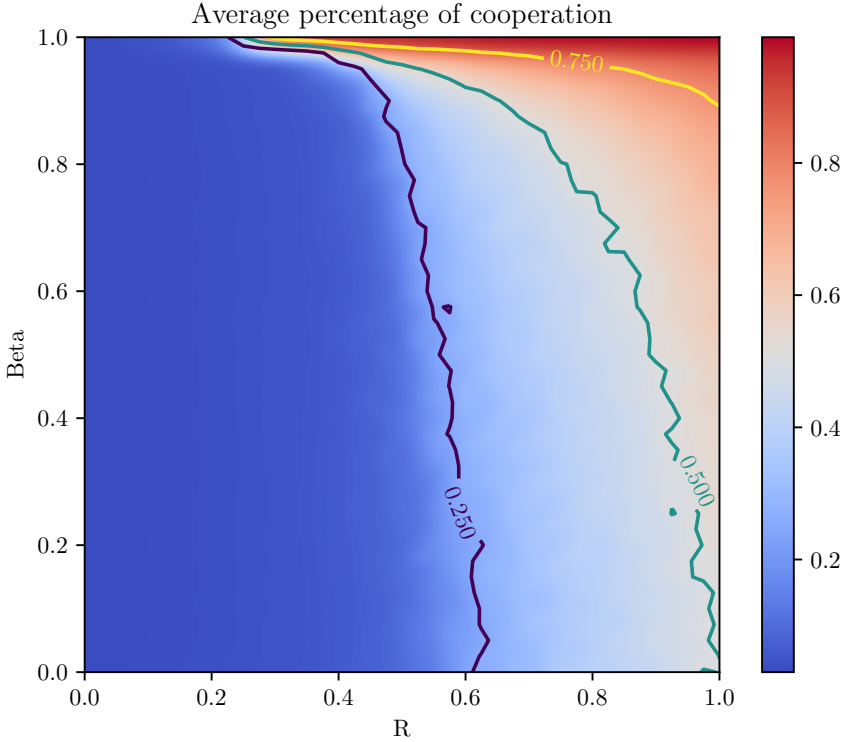


FIGURE 3.4: The figure depicts the average percentage of contributions as a function of the rate of return R and the level of noise in the matching mechanism β . For values of $R \leq \frac{1}{s} = 0.25$ and/or too low values of β , we observe no cooperation. Otherwise, as predicted, the average level of cooperation weakly increases with increasing R and β , and we observe almost complete cooperation when approaching one. The contour lines indicate when the average contribution is above 25%, 50% and 75% of the whole population, respectively. The simulation was obtained for the following set of parameters: $N = 80$, $s = 4$, $\lambda = 30$, $EN = 100$, $p = 0.99$, $T = 2000$.

3.6 SUMMARY

Our analysis aimed to achieve three things. First, we study how the existence of high-contributions Nash equilibria depends on the strategy space of the players and on the homogeneity/heterogeneity of their initial en-

dowment Second, we allow noisy/fuzzy implementations of the mechanism everywhere between “no meritocracy” (completely random grouping) and “full meritocracy” (perfect contribution-based grouping). We found that the minimum meritocracy threshold that enables pure-strategy equilibria with high contributions decreases with the population size, the number of groups and with the rate of return. Where they exist, we established that, in the presence of some unconditional contributors, the high-contribution equilibria are selected rather than the non-contribution equilibrium. Mixed-strategy equilibria do not exist, unless the contribution space is binary (or very coarse).

Finally, we assessed the welfare properties of candidate equilibria to identify the welfare-maximizing regime, given varying degrees of inequality aversion of the regime. We found that setting meritocracy at the minimum meritocracy threshold that enables high-contribution equilibria maximizes welfare for general social welfare functionals as the population becomes infinitely large. An exception to this is in the case of coarse/binary action space and heterogeneous endowments where a mechanism with perfect matching is preferred regardless of how inequality averse the social planner is. For finite populations, the same result holds if (a) the inequality aversion is not extreme and (b) the rate of return is high.

More broadly, the mechanism we considered here belongs to a family of “grouping” mechanisms that enable high-contribution equilibria in voluntary contribution games.¹⁵ Their common feature is that they implement non-random, contribution-based grouping, but require no payoff transfers between players. Importantly, these mechanisms incentivize contributions even if most of the players are narrowly self-interested. Our paper therefore complements other research on cooperative phenomena that arise from other-regarding preferences [130, 131], in particular in public goods games (e.g., [132]).¹⁶ We found that a fuzzy mechanism implementation could outperform a perfect implementation. It is an avenue for future research to consider fuzzy implementations of these alternative grouping mechanisms and of other mechanisms and to study other social dilemmas and games more generally under fuzzy mechanism design.

¹⁵ Other mechanisms include: [122, 123] consider endogenous group formation via voting; [124, 125] study free group entry and exit; [126] analyze roommate-problem stable matching in pairwise-generated public goods; [127] study rematching based on reputation; [128, 129] consider signaling.

¹⁶ Using the terminology of [133], our paper therefore studies a ‘system’ rather than moral ‘acts’ or ‘intentions’. In our mechanism, the system assortments contributions, i.e., actions, as other evolutionary biology mechanisms that lead to cooperation as, for example, kin selection (e.g., [134–136]), local interaction and/or assortative matching of preferences [137–141].

3.7 APPENDIX: PURE STRATEGY NASH EQUILIBRIA

3.7.1 *Perfect Meritocracy*

Let us consider the case of fully-heterogeneous agents playing the binary version of the voluntary contribution game.

In this case, we have $c_i \in \{0, 1\}$, and we define $\mathbf{c} = \{c_i\}_{i \in N}$

We indicate with w_i the endowment of player i , and (without loss of generality) we order players such that $w_1 > w_2 > \dots > w_n$.

We will assume that the endowments of the players do not differ too much; in particular, we will assume that there exists a real finite number X such that $w_i < w_{i-1} + X \forall i$. If this assumption is not satisfied (or if the X is too big), then the only possible Nash equilibrium (NE) is non-contribution by all.

Formalism:

- We have a population of n players with $N = \{1, 2, \dots, n\}$.
- We have g groups, $\{S_1, S_2, \dots, S_g\}$, of size s .
- $\phi_i(c_i | \mathbf{c}_{-i})$ is the payoff of agent i with $\mathbf{c}_{-i} \equiv \mathbf{c} \setminus \{c_i\}$.

Proposition 3.7.1. $\{c_i\} = 0 \forall i$ is an NE.

Proof. This is trivially proven: if every player but j were to contribute $c_i = 0$, the payoff for j would be the following:

$$\phi_j(c_j | \mathbf{c}_{-j}) = w_j(1 - c_j) + R w_j c_j$$

and for values of R less than one, the first order condition (FOC) results in $c_j = 0$. □

We will now study the conditions under which there can be “high efficient” Nash equilibria.

First of all, we note that a situation where the S wealthiest players contribute and all the others do not is a Nash equilibrium:

Lemma 3.7.2. $\{c_i\} = 1 \forall i \in \{1 \dots S\}$ and $\{c_i\} = 0 \forall i \notin \{1 \dots S\}$ is an NE if the rate of return R is greater than or equal to a threshold value \bar{R} .

Proof. Remark: It is enough to check that Player 1 (i.e., the wealthiest player) does not deviate from the contributive strategy because he/she is the player that benefits the least by contributing one.

The expected payoffs for Player 1 are:

$$E[\phi_1(c_1 = 1)] = R \left(\sum_{i=1}^S w_i \right)$$

and:

$$E[\phi_1(c_1 = 0)] = w_1 + \frac{R}{n-S+1} (w_2 + w_3 + w_4)$$

For $\{c_i\} = 1 \forall i \in \{1 \dots S\}$ to be an NE, we need to have that $E[\phi_1(c_1 = 1)] \geq E[\phi_1(c_1 = 0)]$.

Because of the decreasing wealth in the order of players, we have:

$$E[\phi_1(c_1 = 1)] > sRw_S \equiv A$$

and:

$$E[\phi_1(c_1 = 0)] < w_1 + \frac{(S-1)Rw_S}{n-S+1} \equiv B$$

Thus, it is enough to prove that $A \geq B$.

Since $w_S > w_1 + (S-1)X$, we can rewrite $A \stackrel{?}{\geq} B$ as:

$$SR[w_1 - (S-1)X] - \frac{(S-1)}{n-S+1} [w_1 - (S-1)X] \stackrel{?}{\geq} w_1$$

which is equivalent to:

$$w_1(SR-1) - (S-1) \frac{Rw_1}{n-S+1} + SRX \left(1 - \frac{2}{n-S+1} - S + \frac{S}{n-S+1} \right) + \frac{XR}{n-S+1} \geq 0$$

thus leading to:

$$R \geq \frac{w_1}{SR + SX \left(1 - S + \frac{S-2}{n-S+1} \right) - \frac{(S-1)w_1 + X}{n-S+1}} \equiv \bar{R}$$

□

Remark: Note that for $X = 0$ (i.e., for homogeneous agents), the threshold \bar{R} reduces to the threshold found for homogeneous agents by [38].

We now proceed to show that all strategies (barring the one where everyone contributes) where the $k \cdot S$ (with k an integer number) richest players contribute and the rest defect are Nash equilibria of the game:

Lemma 3.7.3. *With $m < n - S$ and k such that $k \equiv m \pmod{S}$:*

If $\{c_i\} = 1 \forall i \in \{1, \dots, k \cdot S\}$, $c_j = 0$ else, is an NE, then $\{c_i\} = 1 \forall i \in \{1, \dots, (k+1) \cdot S\}$, $c_j = 0$ else, is also an NE.

Remark: Note that $m < n - S$ excludes the situation where all players are contributing.

Proof. Because $w_i > w_{i-1} > \dots$, player kS can never be ranked lower than player kS if contributing. Hence, the remaining $n - kS$ players are playing the same game as in Lemma 3.7.2, but with different numbers of players and groups: $\tilde{n} = n - k \cdot S$ and $\tilde{g} = g - k$.

Hence, the same logic of Lemma 3.7.2 applies with the difference that, because n appears only in the numerator of $\bar{R}(n)$, the threshold $\bar{\bar{R}}$ for the reduced game is smaller than the one for the full game; i.e., $\bar{\bar{R}} < \bar{R}$. \square

In all the equilibria described above, players in the last group left, i.e., the one with the S poorest player when all the other players contribute, will always defect:

Lemma 3.7.4. $c_i = 0 \forall i \in \{k \cdot (S - 1) + 1, \dots, k \cdot S\}$.

Proof. This follows trivially from the fact that with no possibility to be placed in a better group because of their low initial endowment, the poorest S players are playing the within the group best response where it is always true that $E[\phi_i(c_i = 1)] < E[\phi_i(c_i = 0)]$. Because $R < 1$, it is always the case that:

$$w_i + R \sum_{j \neq i} w_j > R \sum_j w_j$$

\square

The above lemmata lead to the conclusion that for binary action space and heterogeneous players, a maximum of g pure strategy Nash equilibria can coexist, depending on the value of the marginal per capita rate of return. Interestingly, for the equilibrium with $k = 1$ to exist, the marginal rate of return has to be higher than for the equilibrium with $k = 2$ to exist, and so on, until the lowest threshold for R such that the only contributive equilibrium is the one where all but one group is filled with contributors.

Proposition 3.7.5. *Depending on the value of the marginal rate of return R , there exists a maximum of $g = \frac{S}{n}$ pure strategy Nash equilibria.*

Their structure is so that the $k \cdot s$ ($0 \leq k < g$) poorest players defect and the $(g - k) \cdot s$ richest players contribute.

The non-contribution equilibrium ($k = 0$) always exists, regardless of the value of the marginal rate of return.

There can be no other pure strategy NE.

Proof. Starting from the NE described in Lemma 3.7.2 and iteratively applying Lemma 3.7.3, we can derive the existence of the $g - 1$ contributive equilibria. Note that the threshold for the marginal rate of return derived in Lemma 3.7.2 is the highest of all the thresholds.

Proposition 3.7.1 shows that the non-contribution equilibrium always exists regardless of the value of R .

The heterogeneity of players' endowment ensures that there can be no mixed group such that contributors and non-contributors are grouped together. This, together with Lemma 3.7.4, proves that there can be no other pure strategy Nash equilibria other than the ones described above. \square

Let us now move to continuous action space and show that, in the case of homogeneous agents, when some strategies in equilibrium have positive contributions, there can be no player i contributing $0 < c_i < 1$.

Lemma 3.7.6. *If in equilibrium there is some player i for which c_i s.t. $c_i > 0$, then there cannot be any player j contributing $0 < c_j < 1$.*

Proof. If there are no players other than player j contributing, then from Proposition 3.7.7, it follows that j 's best reply is to defect, as well.

If there are other players contributing more than zero, let us call j the player(s) with the highest contribution.

If there are no ties regarding group membership, j could reduce his/her contribution to $c_j - \varepsilon$, with ε arbitrarily small, and remain in the same group. If there are ties, j could increase his/her contribution by an arbitrary small amount and be sure to be grouped together with players contributing. Similarly to Lemma 3 in [38], it is possible to prove that for an arbitrary small ε , we have:

$$E [\phi (c_j)] < E [\phi (c_j + \varepsilon)]$$

and hence that $c_j \neq \{0, 1\}$ cannot be an equilibrium strategy. \square

3.7.2 Fuzzy Mechanism

We now turn our attention to Nash equilibria in the case of the implementation of the fuzzy mechanism.

Proposition 3.7.7. *For any population size $n > s$, group size $s > 1$, rate of return $r \in (1, s)$ and meritocratic matching factor $\beta \in [0, 1]$, universal non-contribution is always a Nash equilibrium.*

The proof of Proposition 3.7.7 follows from the fact that, given any β and for c_{-i} such that $\sum_{j \neq i} c_j = 0$, we have:

$$1 = \mathbf{E} [\phi_i(0|c_{-i})] > \mathbf{E} [\phi_i(1|c_{-i})] = R. \quad (3.4)$$

Equation (3.4), in other words, means that it is never the best response to be the only contributor for any level of β . If, for any level of β , given any c_{-i} , $\mathbf{E} [\phi_i(0|c_{-i})] > \mathbf{E} [\phi_i(1|c_{-i})]$ holds for all i , then we have a situation where non-contribution is the strictly dominant strategy. In that case, for any level of meritocracy (β), universal non-contribution is the unique Nash equilibrium.

Lemma 3.7.6 implies that for players with homogeneous endowments, it is enough to focus on the extremal strategies when analyzing Nash equilibria.

We shall now proceed to show that additional high-contribution equilibria exist if the marginal per capita rate of return (R) and the meritocratic matching fidelity (β) are high enough. Before we do this, we write 1^m to refer to a strategy profile where “ m players contribute, all others free-ride” and 1_{-i}^m for the same statement excluding player i , i.e., “ m of the other players contribute, $n - m$ other players free-ride”. We also define a critical threshold for the marginal per capita rate of return of $m\text{pcr} = \frac{n-s+1}{ns-s^2+1}$, which is the threshold, as identified by [38], for which the high-contribution equilibria exist for the case of perfect merit (when $\beta = 1$) and homogeneous endowments.

Proposition 3.7.8. *Given population size $n > s$, group size $s > 1$ and rate of return r such that $R \in (m\text{pcr}, 1)$, there exists a necessary meritocracy level, $\beta \in (0, 1)$, above which there is a high-contribution Nash equilibrium, where $m > 0$ agents contribute and the remaining $n - m$ agents free-ride.*

Proof. The following two conditions must hold for Proposition 3.7.8 to be true:

$$\mathbf{E} [\phi_i(1|1_{-i}^m)] \geq \mathbf{E} [\phi_i(0|1_{-i}^{m-1})] \quad (3.5)$$

$$\mathbf{E} [\phi_i(0|1_{-i}^m)] \geq \mathbf{E} [\phi_i(1|1_{-i}^{m+1})] \quad (3.6)$$

The proof for the existence of an equilibrium in which some appropriate (positive) number of contributors m exists for the case when $\beta = 1$ and

$R \geq mpcr$ follows from Theorem 1 in [38], in which case both Equations (3.5) and (3.6) are strictly satisfied.

The fixed point argument behind that result becomes clear by inspection of Terms (ii) and (iii) in Expression (3.2): namely, the decision to contribute rather than to free-ride is a trade-off between (ii) ‘the sure loss on own contribution’, which is zero for free-riding, versus (iii) ‘the expected return on others’ contributions’, which may be larger by contributing rather than by free-riding depending on how many others also contribute. Obviously, when c_{-i} is such that $\sum_{j \neq i} c_j = 0$ or $\sum_{j \neq i} c_j = (n-1)$ (i.e., if either all others free-ride or all others contribute), it is the case that $\phi_i(0|c_{-i}) > \phi_i(1|c_{-i})$. Hence, in equilibrium, $0 < m < n$.

Now, suppose 1^m describes a pure-strategy Nash equilibrium for $\beta = 1$ with $0 < m < n$ and $R \in (mpcr, 1)$ in which case Equations (3.5) and (3.6) are strictly satisfied. Note that β has a positive effect on the expected payoff of contributing and a negative effect on the expected payoff of free-riding:

$$\partial E [\phi_i(1|1^m_{-i})] / \partial \beta > 0 \quad (3.7)$$

$$\partial E [\phi_i(0|1^m_{-i})] / \partial \beta < 0 \quad (3.8)$$

When $\beta = 0$, we know that $\phi_i(1|1^m_{-i}) = R < \phi_i(0|1^m_{-i}) = 1$ for any m . However, by existence of the equilibrium with $m > 0$ contributors when $\beta = 1$, provided that $R > mpcr$ is satisfied, there must exist some maximum value of $\underline{\beta} \in (0, 1)$, at which either Equation (3.5) or Equation (3.6) first binds due to continuity of Expressions (3.7) and (3.8) in β . That level is the bound on β above which the pure-strategy Nash equilibrium with $m > 0$ exists. \square

Remark 3.7.9. *Note that, for a finite population of size n , a group size s larger than one implies that $mpcr > 1/s$ for Proposition 3.7.8 to be true, but as $n \rightarrow \infty$, $mpcr$ converges to $1/s$.¹⁷*

A special case of such a pure-strategy Nash equilibrium is the high-contribution Nash equilibrium as identified by (see [38]): in our setup, the almost-no-free-riders pure-strategy Nash equilibrium generalizes to the pure-strategy Nash equilibrium in which m is chosen to be the largest value given n, s, r for which Equations (3.5) and (3.6) hold. For that m to be larger than zero, β needs to be larger than $\underline{\beta}$ (Proposition (3.7.8)).

¹⁷ It is easy to check that $\lim_{n \rightarrow \infty} mpcr = 1/s$.

3.8 APPENDIX: MIXED-STRATEGY NASH EQUILIBRIA

Now, we shall compare the asymmetric high-contribution Nash equilibria with potential symmetric mixed-strategy Nash equilibria. For this, we define $p_i \in [0, 1]$ as a mixed strategy with which player i plays ‘contributing’ ($c_i = 1$), while playing ‘free-riding’ ($c_i = 0$) with $(1 - p_i)$. Write $p = \{p_i\}_{i \in N}$ for a vector of mixed strategies. Write 1^p for “all players play p ” and 1_{-i}^p for the same statement excluding some player i .

First, we consider the case of a binary contribution space, i.e., when contributions can only be full or null.

Proposition 3.8.1. *Consider the case of $c_i \in \{0, 1\}$ for all i . Given population size $n > s$ and group size $s > 1$, there exists a rate of return r such that $R \in [mpcr, 1)$ beyond which there exists a **necessary meritocracy** level, $\underline{\beta} \in (0, 1)$, such that there always are two mixed strategy profiles, where every agent places weight $p > 0$ on contributing and $1 - p$ on free-riding, that constitute a **symmetric mixed-strategy Nash equilibrium**. One will have a high \bar{p} (the near-efficient symmetric mixed-strategy Nash equilibrium), and one will have a low p (the less-efficient symmetric mixed-strategy Nash equilibrium).*

Proof. The symmetric mixed-strategy Nash equilibrium exists if there exists a $p \in (0, 1)$ such that, for any i ,

$$\mathbf{E} \left[\phi_i(0|1_{-i}^p) \right] = \mathbf{E} \left[\phi_i(1|1_{-i}^p) \right], \quad (3.9)$$

because, in that case, player i has the best response also playing $p_i = p$, guaranteeing that 1^p is a Nash equilibrium. Proposition 3.7.8 implies that, if $R > mpcr$, Equations (3.5) and (3.6) are strictly satisfied when $\beta = 1$ for m contributors corresponding to the almost-no-free-riders pure-strategy Nash equilibrium. Indeed, Expressions (3.5) and (3.6) imply lower and upper bounds (see [38]) on the number of free-riders given by:

$$l = \frac{n - nR}{1 - R + nR - r}, \quad u = 1 + \frac{n - nR}{1 - R + nR - r}. \quad (3.10)$$

Part 1. First, we will show, given any game with population size n and group size s , for the case when $\beta = 1$, that there is (i) at least one symmetric mixed-strategy Nash equilibrium when $R \rightarrow 1$; (ii) possibly none when $R = mpcr$; and (iii) a continuity in R such that there is some intermediate value of $R \in [mpcr, 1)$ above which at least one symmetric mixed-strategy Nash equilibrium exists, but not below.

(i) Because $\partial \mathbf{E} \left[\phi_i(c_i | 1_{-i}^p) \right] / \partial p > 0$ for all c_i , there exists a $p \in (\frac{m-1}{n}, \frac{m+1}{n})$ such that Expression (3.9) holds if $R \rightarrow 1$. This is the standard symmetric mixed-strategy Nash equilibrium, which always exists in a symmetric two-action n -person game where the only pure-strategy equilibria are asymmetric and of the same kind as the high-contribution pure-strategy Nash equilibrium (see the proof of Theorem 1 in [142]). In this case, the presence of the non-contribution Nash equilibrium makes no difference because the incentive to free-ride vanishes as $R \rightarrow 1$.

(ii) If $R = mpcr$, one or both of the equations, (3.5) or (3.6), bind. Hence, unless Expression (3.9) holds exactly at $p = m/n$ (which is a limiting case in n that we will address in Proposition 3.8.4), there may not exist any p such that Expression (3.9) holds. This is because the binomially-distributed proportions of contributors implied by p , relatively speaking, place more weight on the incentive to free-ride than to contribute because universal free-riding is consistent with the non-contribution Nash equilibrium, while universal contributing is not a Nash equilibrium. In this case, the incentive to free-ride is too large for a symmetric mixed-strategy Nash equilibrium to exist.

(iii) $\partial \mathbf{E} \left[\phi_i(c_i | 1_{-i}^p) \right] / \partial r$ is a different linear, positive constant for both $c_i = 0$ and $c_i = 1$. At and above some intermediate value of R , therefore, there exists a $p \in (0, 1)$ such that, if played in a symmetric mixed-strategy Nash equilibrium, the incentive to free-ride is mitigated sufficiently to establish Equation (3.9). We shall refer to this implicit minimum value of R by \overline{mpcr} .

Finally, for any $p > 0$ constituting a symmetric mixed-strategy Nash equilibrium when $\beta = 1$, $\mathbf{E} \left[\phi_i(0 | 1_{-i}^p) \right] = \mathbf{E} \left[\phi_i(1 | 1_{-i}^p) \right] > 1$. Because of this, a similar argument as in Proposition 3.7.8 applies to ensure the existence of some $\underline{\beta} \in (0, 1)$ above which the symmetric mixed-strategy Nash equilibrium continues to exist when $R > \overline{mpcr}$: because, at $\beta = 1$, Equations (3.5) and (3.6) are strictly satisfied and $\mathbf{E} \left[\phi_i(0 | 1_{-i}^p) \right] = \mathbf{E} \left[\phi_i(1 | 1_{-i}^p) \right] > 1$, there therefore must exist some $\beta < 1$ and $p' < p$ satisfying Equation (3.9) while still satisfying $\mathbf{E} \left[\phi_i(0 | 1_{-i}^p) \right] = \mathbf{E} \left[\phi_i(1 | 1_{-i}^p) \right] > 1$. Note that this implicit bound here may be different from that in Proposition 3.7.8.

Part 2. If $R > \overline{mpcr}$ and $\beta > \underline{\beta}$, the existence of two equilibria with $\bar{p} > p > 0$ is shown by analysis of the comparative statics of Equation (3.9).

First note that, for any $R > \overline{mpcr}$ and $\beta > \underline{\beta}$, $\partial \mathbf{E} \left[\phi_i(0 | 1_{-i}^p) \right] / \partial \beta < 0$, while $\partial \mathbf{E} \left[\phi_i(1 | 1_{-i}^p) \right] / \partial \beta > 0$. p therefore has to take different values for

Equation (3.9) to hold for two different values of β above $\underline{\beta}$. It is unclear whether it has to take a higher or lower value. Note also that both $\partial \mathbf{E} [\phi_i(0|1^p_{-i})] / \partial p > 0$ and $\partial \mathbf{E} [\phi_i(1|1^p_{-i})] / \partial p > 0$ for all $\beta \in (0, 1)$. We can rearrange the partial derivative with respect to β of Expression 3.9, and obtain:

$$\partial p / \partial \beta = \frac{\partial \mathbf{E} [\phi_i(1|1^p_{-i})] / \partial \beta - \partial \mathbf{E} [\phi_i(0|1^p_{-i})] / \partial \beta}{\partial \mathbf{E} [\phi_i(0|1^p_{-i})] / \partial p - \partial \mathbf{E} [\phi_i(1|1^p_{-i})] / \partial p}. \quad (3.11)$$

Expression 3.11 is negative if the denominator is negative, because the numerator is always positive.

Claim 3.8.2. *The denominator of Equation (3.11) is negative when p is low, and positive when p is high.*

Write $\bar{w}_{c_i}^i$ and $w_{c_i}^i$ respectively for the probabilities with which agent i is matched in an above- or below-average group when playing c_i where the average is taken over contributions excluding i . Write $\mathbf{E} [\bar{\phi}_i(c_i|1^p_{-i})]$ and $\mathbf{E} [\phi_i(c_i|1^p_{-i})]$ for the corresponding expected payoffs.

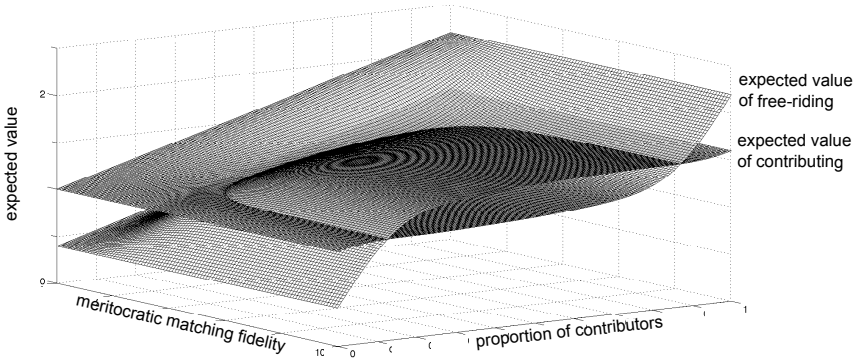


FIGURE 3.5: The figure depicts the expected payoffs of contributing versus free-riding for the economy with $n = 16$, $s = 4$, $r = 1.6$. The expected values of $\phi_i(0|1_{-i}^p)$ and $\phi_i(1|1_{-i}^p)$ are plotted as functions of contribution probability p and meritocratic matching fidelity β . The two planes intersect at the bifurcating symmetric mixed-strategy Nash equilibrium-values of \bar{p} and p (see Proposition 3.8.1). Notice that the expected values of both actions increase linearly in p when the meritocratic matching fidelity is zero, but turn increasingly S-shaped for larger values, until they intersect at \bar{p} and p .

Recall that, for $\beta > 0$ and $1_{-i}^p \in (0, 1)$, Expression (3.15)¹⁸ holds, where \hat{k} is compatible with a perfect ordering $\hat{\pi}$, and \tilde{k} is any rank compatible with a mixed ordering $\tilde{\pi}$. When $1_{-i}^p = 0$ or $1_{-i}^p = 1$, the probability of agent i to take rank j , f_{ij}^β , depends on his/her choice of c_i , but $\bar{w}_{c_i}^i = w_{c_i}^i = 0$ for any choice of contribution c_i .

For $p \in (0, 1)$, we shall rewrite $\partial \mathbf{E} [\phi_i(0|1_{-i}^p)] / \partial p$ in the denominator of Equation (3.11) as:

$$\frac{\partial}{\partial p} [\bar{w}_0^i * \mathbf{E} [\bar{\phi}_i(0|1_{-i}^p)]] + \frac{\partial}{\partial p} [w_0^i * \mathbf{E} [\phi_i(0|1_{-i}^p)]] \quad (3.12)$$

and $\partial \mathbf{E} [\phi_i(1|1_{-i}^p)] / \partial p$ as:

$$\frac{\partial}{\partial p} [\bar{w}_1^i * \mathbf{E} [\bar{\phi}_i(1|1_{-i}^p)]] + \frac{\partial}{\partial p} [w_1^i * \mathbf{E} [\phi_i(1|1_{-i}^p)]] . \quad (3.13)$$

¹⁸ See Appendix C.

Notice that, for large β , $w_0^i \gg \bar{w}_0^i$ when p is close to zero, and $w_1^i \ll \bar{w}_1^i$ when p is close to one. Moreover, notice that the existence of the pure-strategy Nash equilibrium with high contribution for high levels of β ensures that $\mathbf{E}[\phi_i(0|1_{-i}^p)]$ is not always larger than $\mathbf{E}[\phi_i(1|1_{-i}^p)]$. It therefore follows from continuity in β that Expression 3.13 exceeds Expression 3.12 when p is low and that Expression 3.12 exceeds Expression 3.13 when p is high; hence, the denominator of Equation 3.11 is negative when p is low and positive when p is high. Figure 3.6 illustrates. \square

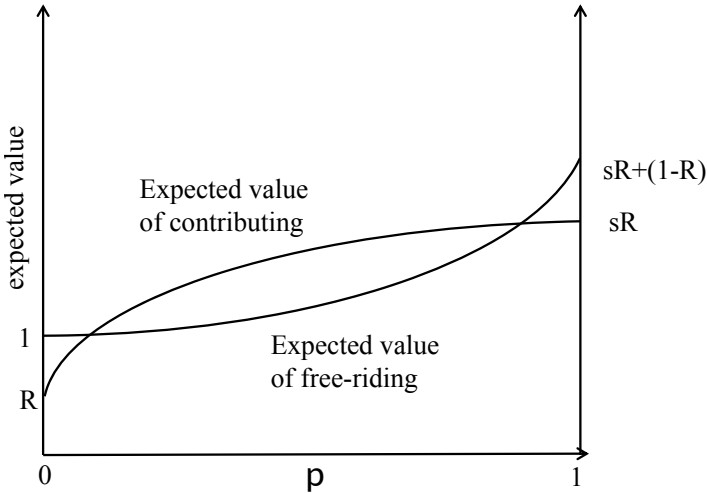


FIGURE 3.6: Expected payoffs of contributing versus free-riding. The expected values of $\phi_i(0|1_{-i}^p)$ and $\phi_i(1|1_{-i}^p)$ are plotted as functions of the probability p for some fixed $\beta > \underline{\beta}$. The two planes intersect at the bifurcating symmetric mixed-strategy Nash equilibrium-values of \bar{p} and p (see Proposition 3.8.1). The relative slopes of the two curves illustrate the proposition. Note that this figure is a slice through Figure 3.5 along a value of $\beta > \underline{\beta}$.

Remark 3.8.3. Note that the necessary meritocracy level $\underline{\beta}$ in Propositions 3.7.8 and 3.8.1 need not be the same. We shall write $\underline{\beta}$ for whichever level is larger.

Corollary 3.8.3.1. For intermediate values of p , contributing is a better reply than free-riding.

Corollary 3.8.3.2. For the case when $c_i \in [0, 1]$, a symmetric mixed-strategy Nash equilibrium does not exist.

Proof. Consider the symmetric mixed-strategy Nash equilibrium from the above proof with $p_i = p$ for all i . Suppose all $j \neq i$ play p . Then, in the neighborhood of $\epsilon = 0$, $\partial \mathbf{E} [\phi_i(\epsilon | 1_{-i}^p)] / \partial \epsilon > 0$, because playing $\epsilon > 0$ for this player will always rank him above others playing zero. Hence, $p_i = p$ for all i cannot be an equilibrium. \square

Proposition 3.8.4. *Given group size $s > 1$, then, if $\beta = 1$, as $n \rightarrow \infty$: (i) $1^m/n$ of the high-contribution pure-strategy Nash equilibrium and \bar{p} of the near-efficient symmetric mixed-strategy Nash equilibrium converge; and (ii) the range of R for which these equilibria exist converges to $(1/s, 1)$.*

Proof. Suppose $R > \overline{mp\bar{c}v}$, i.e., that both symmetric mixed-strategy Nash equilibrium and high-contribution pure-strategy Nash equilibrium exist. Let 1^m describe the high-contribution pure-strategy Nash equilibrium and 1^p describe the near-efficient symmetric mixed-strategy Nash equilibrium. Recall that expressions under (3.10) summarize the lower and upper bound on the number of free-riders, $(n - m)$ in the high-contribution pure-strategy Nash equilibrium. Taking $\lim_{n \rightarrow \infty}$ for those bounds implies a limit lower bound of $\frac{1}{1+n\frac{R-r/n}{1-R}}$ and a limit upper bound of the expected proportion of free-riders of $\frac{1}{n} + \frac{1}{1+n\frac{R-r/n}{1-R}}$ and, thus, bounds on the number of free-riders that contain at most two integers and at least one free-rider (notice that the limits imply that exactly one person free-rides as $R \rightarrow 1$). We know that, if there is one more free-rider than given by the upper bound, then Equation (3.6) is violated. Similarly, if there is one fewer free-rider than given by the lower bound, then Equation (3.5) is violated.

With respect to the near-efficient symmetric mixed-strategy Nash equilibrium, recall that Expression 3.9 must hold; i.e., $\mathbf{E} [\phi_i(0 | 1_{-i}^p)] = \mathbf{E} [\phi_i(1 | 1_{-i}^p)]$. We can rewrite $\mathbf{E} [\phi_i(c_i | 1_{-i}^p)]$ as $\mathbf{E} [\phi_i(c_i | B)]$, where B is the number of other players actually contributing (playing $c_i = 1$), which is distributed according to a binomial distribution $\text{Bin}(p, n)$ with mean $\mathbf{E}[B] = np$ and variance $\mathbf{V}[B] = np(1-p)$. As $n \rightarrow \infty$, by the law of large numbers, we can use the same bounds obtained for the high-contribution pure-strategy Nash equilibrium to bound $(B/n) \in [(n-u)/n, (n-l)/n]$, which converges to the unique p for which Expression (3.9) actually holds.¹⁹

Suppose all players contribute with probability p corresponding to the near-efficient symmetric mixed-strategy Nash equilibrium limit value. Then,

¹⁹ Details concerning the use of the law of large numbers can be followed based on the proof in [142].

$\lim_{n \rightarrow \infty} \mathbf{V}[(B/n)] = \lim_{n \rightarrow \infty} \frac{p(1-p)}{n} = 0$ for the actual proportion of contributors. Hence, the limit for the range over R necessary to ensure existence converges to that of the high-contribution pure-strategy Nash equilibrium, which by Remark 3.7.9 is $(1/s, 1)$. \square

Remark 3.8.5. *In light of the limit behavior, it is easy to verify, ceteris paribus, that the value of the marginal per capita rate of return necessary to ensure the existence of the symmetric mixed-strategy Nash equilibrium is decreasing in population size n , but increasing in group size s ; i.e., decreasing in relative group size s/n .*

3.9 APPENDIX: PROPERTIES OF THE FUZZY RANKING

Fixing $\beta \in [0, 1]$, we indicate with $\bar{k}_i^\beta(c_i)$ the expected ranking of player i when contributing c_i . The ranking is such that, under a perfect ordering, $k_i > k_j \Rightarrow c_i \leq c_j$. We define the fuzziness such that for each player, the observed contribution x_i is distributed normally around the true contribution c_i with standard deviation σ . We define $\sigma^2 = \frac{1-\beta}{\beta}$ such that $x_i \sim \mathcal{N}(c_i, \sigma^2)$.

The following properties hold:

- For $\beta = 1$, the contributions are perfectly observable, and the ranking follows a perfect ordering.
- For $\beta = 0$, contributions do not matter, and the ordering of the players is determined completely at random. This follows from the property of the normal distribution for $\sigma \rightarrow \infty$ ($\beta \rightarrow 0$).
- For $0 < \beta < 1$, the expected ranking for every player has the following properties:

$$\frac{\partial \mathbf{E} \left[\bar{k}_i^\beta(c_i) \right]}{\partial c_i} < 0 \quad (3.14)$$

$$\mathbf{E} \left[\bar{k}_i^\beta(c_L) - \bar{k}_i^\beta(c_H) \right] > 0 \quad ; \quad c_L < c_H \quad (3.15)$$

$$\frac{\partial \mathbf{E} \left[\bar{k}_i^\beta(c_L) - \bar{k}_i^\beta(c_H) \right]}{\partial \beta} > 0 \quad (3.16)$$

Proofs

Let us define x_L and x_H as the imperfectly observed contributions when a player contributes c_L and c_H , respectively; i.e., $x_L \sim \mathcal{N}(c_L, \sigma^2)$ and $x_H \sim \mathcal{N}(c_H, \sigma^2)$.

Lemma 3.9.1. *Proof of (3.14):*

Proof. This follows trivially from $\mathbf{E} \left[\bar{k}_i^\beta(c_q) \right] = \bar{k}_i^\beta(\mathbf{E}[c_q])$ and the matching rules. \square

Lemma 3.9.2. *Proof of (3.15):*

Proof. It again follows from $\mathbf{E} \left[\bar{k}_i^\beta(c_q) \right] = \bar{k}_i^\beta(\mathbf{E}[c_q])$, and hence:

$\mathbf{E} \left[\bar{k}_i^\beta(c_L) - \bar{k}_i^\beta(c_H) \right] = \mathbf{E} \left[\bar{k}_i^\beta(c_L) \right] - \mathbf{E} \left[\bar{k}_i^\beta(c_H) \right] = \bar{k}_i^1(c_L) - \bar{k}_i^1(c_H) > 0$
for $c_L < c_H$. \square

Lemma 3.9.3. *Proof of (3.16):*

Proof. Let us write Equation (3.16) as a function of σ :

$$\frac{\partial g}{\partial \beta} = -\frac{1}{\beta^2} \frac{\partial g}{\partial \sigma^2} = -\frac{1}{2\sigma\beta^2} \frac{\partial g}{\partial \sigma}$$

with a slight abuse of notation. Hence, we have that:

$$\frac{\partial \mathbf{E} \left[\bar{k}_i^\beta(c_L) - \bar{k}_i^\beta(c_H) \right]}{\partial \beta} > 0 \Rightarrow \frac{\partial \mathbf{E} \left[\bar{k}_i^\beta(c_L) - \bar{k}_i^\beta(c_H) \right]}{\partial \sigma} < 0.$$

The expected difference in ranking between c_L and c_H can be rewritten in terms of the probability of observing a higher realized contribution for c_L than for c_H ; i.e.,

$$\frac{\partial \mathbf{E} \left[\bar{k}_i^\beta(c_L) - \bar{k}_i^\beta(c_H) \right]}{\partial \sigma} < 0 \Leftrightarrow \frac{\partial P(z > 0)}{\partial \sigma} < 0$$

with $z \equiv x_L - x_H$. Note that $z \sim \mathcal{N}(c_L - c_H, 2\sigma^2)$ ²⁰.

By definition, we have:

$$P(z > 0) = 1 - P(z \leq 0) \equiv 1 - \Phi \left(\frac{c_H - c_L}{\sqrt{2}\sigma} \right) = \frac{1}{2} \left[1 - \operatorname{erf} \left(\frac{c_H - c_L}{2\sigma} \right) \right]$$

²⁰ See, e.g., [143]

where $\Phi(x)$ denotes the cumulative distribution function of the normal distribution and $erf(x)$ the error function.

We thus obtain:

$$\frac{\partial \mathbf{E} \left[\bar{k}_i^\beta(c_L) - \bar{k}_i^\beta(c_H) \right]}{\partial \beta} > 0 \Leftrightarrow \frac{\partial P(z > 0)}{\partial \sigma} < 0 \Leftrightarrow \frac{\partial erf\left(\frac{c_H - c_L}{2\sigma}\right)}{\partial \sigma} > 0$$

However, $erf(x) = \int_0^x e^{-t^2} dt$, and hence, $\frac{\partial erf(x)}{\partial x} = e^{-x^2}$. We thus obtain that:

$$\frac{\partial erf\left(\frac{c_H - c_L}{2\sigma}\right)}{\partial \sigma} = e^{-\left(\frac{c_H - c_L}{2\sigma}\right)^2} \cdot 2 \left(\frac{c_H - c_L}{2\sigma}\right) \cdot \frac{c_H - c_L}{2\sigma^2} > 0$$

which is always positive for $\beta > 0$. □

HETEROGENOUS AGENTS IN VOLUNTARY CONTRIBUTION GAMES WITH ASSORTATIVE MATCHING AND WEALTH ACCUMULATION

ABSTRACT

Voluntary contribution games (VCG) are a classic example of social dilemma where the dominant strategy for individuals conflicts with what is best for society as a whole. Assortative matching (i.e. the practice of ranking agents based on their contributions and then grouping them according to this ranking) has been shown to be a mechanism leading to desirable (highly cooperative) equilibria in VCGs, due to agents contributing more so that they are grouped with other agents doing the same. However the outcome of the game can change quite drastically if it is repeated over time and heterogeneity among agents is allowed. In this chapter, using an agent based model, we study how VCGs with Assortative Matching evolve when agents can accumulate wealth over time. Several ranking mechanisms are proposed that take into account the possible heterogeneity in wealth and talent among agents and, using a computational approach, we determine which mechanisms are sub-optimal from a social welfare point of view.

4.1 INTRODUCTION

How can one sustain cooperative (or altruistic) behavior in ‘social dilemma’ situations? This has been a recurrent question attracting scholars from several disciplines [73, 144–146]. The dilemma can be summarized with the so called free-riders problem: while a collective benefits from the presence of a public good produced at some cost by cooperative individuals, free-riders can still benefit from the public good without producing any of it, thus having little incentive to pay its cost in the first place. This misalignment of collective interest and individual strategic incentives often results in the so called ‘tragedy of the commons’, i.e. underproduction of the public good [8, 30, 92].

Voluntary contribution games (VCG) [91] are a classic example of social dilemmas: Individuals are placed in one or more groups and asked how much of their starting endowment they would like to contribute to a group-wide common pool; they then keep what they did not contribute and share equally their group output (i.e. the sum of all the contributions made by the members of the group multiplied by the marginal-per-capita-rate-of-return or *mpcr*). When the *mpcr* is bigger than the group size, the total welfare is maximized if every player contributes all of its endowment. However, if players are assigned to the groups in a random fashion, the only dominant strategy for each player is to contribute nothing (for any *mpcr* bigger than one) [20, 21], thus realizing the tragedy of the commons.

Conversely, if the matching of players to groups is linked to the individuals’ contributions decisions, the outcome of the game might change dramatically. Indeed, Gunnthorsdottir et al. [38] showed that when players are asked to pre-commit on how much they are willing to contribute and then groups are formed based on the contributions so that high (low) contributors are matched with other high (low) contributors, there exists an equilibrium in which almost all players contribute everything to the common pool of their groups, thus realizing a near-efficient outcome. This family of mechanisms has been named “meritocratic group-based matching” or “assortative matching”.

Recent scholarship [39] and the previous chapters have tested how robust the positive predictions arising from assortative matching are to variations in the payoff function, the action space and, more importantly, to ex-ante inequality among the players¹.

¹ Allowing for heterogeneity in player’s initial budget.

However, an important limitation of prior studies has been the focus on one-shot games instead of considering more realistic repeated interactions. Moving from one to repeated interactions might substantially alter the equilibria of the game [22, 147, 148], especially if one allows for wealth to be accumulated over time. Indeed, wealth accumulation could have a significant impact on the outcome of the game because it might increase the heterogeneity of the population over time, which could alter the level of cooperative behaviour (see e.g. [96]). Hence, it is crucial to address the question of how does taking into account repeated interactions affect the positive predictions obtained with one-shot VCGs with assortative matching.

As a first step toward dealing with this question more generally, this paper investigates a repeated VCG with assortative matching in which heterogeneous agents, interacting according to a logit-response dynamics, accumulate wealth over multiple rounds (thus resulting in agents having different amount with which to contribute to the common pool). Two diverse dimensions over which agents can differ are taken into account: wealth and talent. This is to model the fact that e.g., workers might have different innate talent no matter how hard they work or investors might have different abilities to judge a good investment and/or start with a different initial capital. Note that, while in a one-shot game the two different sources of heterogeneity would produce the same results, in a repeated game with payoff accumulation they might have different effects.

Building on Gunnthorsdottir [38], agents pre-commit to a certain contribution, are ranked based on it and are then assigned to the groups based on this ranking. However, with the introduction of different sources of agents' diversity, one has automatically introduced different criteria based on which one could assign agents to groups. For example, one could rank the players based on the total contribution that they made to the group or based solely on the percentage of their endowment they contributed, regardless of their talent (i.e. a multiplicative factor of their contribution).

Using an agent based simulation, this paper investigates the model described above for four different ranking criteria² and the results summarize as follows:

The average level of cooperation in the agents' population strongly depends on which criteria it is used to rank the agents. Furthermore, ranking criteria that result in high level of cooperation in the short run, might not sustain it over longer periods of time. In particular, it is found that ranking

² Described in the Model section

agents taking into account their talent, but not their total amount of wealth, is the criterion that leads to the highest levels of cooperation in the long run. The distribution of wealth among the population is also a function of which criterion is used to assign agents to groups: ranking agents only based on their relative contribution to the common pool being results in the most equal distribution of wealth in the population.

A social planner trying to minimize the (wealth) inequality in the population while maximizing society's output, can thus judge which ranking criteria among the ones considered could best serve his purpose. By considering the long term outcome of the different simulated VCGs it is possible to determine that some criteria are worse than others, by e.g. having the same global output as another ranking but distributing it more unevenly. Naturally, which among the remaining ranking criteria should be considered the best, will depend on the social planner's preferences.

The rest of the chapter is structured as follows. In the following section, we set up the model. Section 4.3 contains the chapter's results and section 4.4 concludes. A Methods section contains details of the computational algorithm.

4.2 THE MODEL

Using an agent based model, we simulate the game below in which agents update their strategy using a logit-response dynamics [81, 82, 121]³:

N agents are assigned an initial endowment w_i , the same for every player i , and are randomly assigned a talent r_i drawn from a Normal distribution⁴.

After the initialization, the agents play the following game for T rounds:

1. *Choose strategy.* Agents make a simultaneous and committed unilateral decision (updating their strategy using a logit-response function) regarding how much of their endowment to contribute to a common pool. $\alpha_i \in [0, 1]$ indicates the percentage of w_i contributed by player i .
2. *Groups creation.* Agents are ranked from highest to lowest based on one of four possible criteria (see the ranking subsection below), with

³ The reader is referred to the methods section for a description of the pseudo algorithm used in the simulations.

⁴ Test simulations were run also with different talent distributions but the qualitative outcome of the game did not change.

ties broken at random. Based on the ranking, they are assigned to $M = \frac{N}{S}$ equal-sized groups of size S , such that the S highest-ranking agents are assigned to the first group, the S second-highest ranking players are assigned to the second, etc.

3. *Payoffs.* Every agent receives a payoff P_i realized based on the total contribution in each group and on his decision. Each agent receives the sum of all the contributions made by the members of his group multiplied by their individual talent and by a common factor Q . In formula:

$$P_i = Q \sum_{j \in G_i} \alpha_j w_j r_j$$

with G_i being the group to which agent i belongs.

4. *Wealth update.* Agents add their payoff for the round to the amount of wealth that they did not contribute to the common pool and then play a new round of the game. Hence their updated wealth $w_{NEW} = \phi_i(\alpha_i)$ is computed as:

$$\phi_i(\alpha_i) = w_i(1 - \alpha_i) + P_i = w_i(1 - \alpha_i) + Q \sum_{j \in G_i} \alpha_j w_j r_j \quad (4.1)$$

Note that every round, agents can contribute any percentage of their entire wealth; hence, even though every agent starts with the same initial endowment, after the first round everyone could have a different amount with which to contribute to the common pool⁵.

The *marginal per capita rate of return* Q represents the benefits of cooperation among members of the same group. If agents were to be assigned in groups completely randomly, the above game would be a social dilemma for $\frac{1}{5} < Q < 1$. Indeed, for these intermediate values of Q the group output is maximized if every agent contributes his entire endowment but every agent's best response is to contribute nothing to the common pool.

In contrast, the *talent* r_i is a multiplicative factor that represents how much each individual agent i is able to contribute to the common group. A very low talented agents provides little benefit to his group when contributing to the common pool, while a highly talented agent is highly beneficial to his group when contributing. Talents are drawn so that the above game is still a social dilemma⁶

⁵ In particular, note that players can also lose wealth during a given round.

⁶ See the methods section for more details.

In the game described above, there are two possible sources of heterogeneity:

- *Heterogeneity in wealth*: Agents might end up with different endowments and so the maximum amount of what they can contribute to the common pool might vary. In particular, depending on the ranking criterion, even by contributing everything it has, a poor agent might never be able to enter in one of the top groups, i.e. the groups with the highest return from the common pool.
- *Heterogeneity in talent*: Contributions of different agents might have different effectiveness when added to the common pool, thus resulting in a different marginal per capita rate of return for each agent. Two players contributing the same amount, hence, might be assigned to completely different groups.

4.2.1 Ranking criteria

When assigning agents to their groups, taking or not into account the diversity of the agents might result in completely different outcomes. Based on the above diversity, the ranking criteria that can be used to assign agents to their groups are the following:

1. *Total contribution*: Players are ranked based on their total contribution to the group. This means that they are ranked based on $\alpha_i \cdot w_i \cdot t_i$.
2. *Total wealth*: Players are ranked based on the total wealth they bring to the group, regardless of their talent. This means that they are ranked based on $\alpha_i \cdot w_i$.
3. *Percentage and talent*: Players are ranked based on the percentage of their wealth contributed but taking into account their talent, regardless of their wealth. This means that they are ranked based on $\alpha_i \cdot t_i$.
4. *Percentage of wealth contributed*: Players are ranked solely based on the percentage α_i of their wealth that they contributed.

Each ranking system could be interpreted as mirroring a society designed by social planners with different preferences. E.g. the ranking systems 1 and 2 might be considered to represent two different kinds of "capitalistic" society while the ranking system 4 a society mainly focused on egalitarianism.

4.3 RESULTS

Figure 4.1 shows snapshots of sample realizations of the simulation described above for three different ranking criteria: players ranked based on their total contribution (case 1), players ranked based on the percentage of their contribution taking into account their talent (case 3), and players ranked solely based on the percentage of wealth they contributed (case 4). The case of the players ranked solely based on their total wealth contributed (case 2) is qualitatively similar to 1 and it is therefore not shown⁷ Each subfigure depicts the entire population, with each agent represented by a colored dot, at an earlier (left column) and later (right column) stage of the simulation. The abscissa of an agent indicates which percentage of his wealth he contributed to the common pool while his ordinate indicates his current wealth. The black cross marks the average of the population. The color of each agent indicates how talented he is, ranging from low talent (blue) to high talent (red).

When agents are ranked only based on the percentage of their wealth contributed to the common pool α_i (case 4), the results are in line with what expected from previous scholarship [38, 40] and the previous chapters (see fig. 4.1 (e) and (f)): agents are split in a (relatively) high contributions majority and a low contributions minority. In fact, this is very close to one of the possible equilibria of the one-shot⁸ version of the Voluntary Contribution Game with assortative matching in case of homogeneous agents; in that case, the almost Pareto optimal equilibrium is for nearly all agents to contribute their entire endowment and for few (less than the group size S) to contribute nothing (see e.g. Gunnthorsdottir et al. [38]). This situation seems to remain qualitatively the same throughout the simulation from an early stage (fig. 4.1 (e)) to late stage (fig. 4.1 (f)) Naturally, agents with higher talent tend to have a higher wealth than less talented agents, due to the fact that their own contributions to the common pool make their groups better than average. Nevertheless, the total amount of wealth in the population is somewhat evenly distributed among the agents⁹.

In cases 1 and 3, where agents are ranked taking into account more than only α_i (fig. 4.1 (a), (b),(c), (d)), one can observe that the most talented individuals tend to be wealthier and, more importantly, to contribute less than the other agents. This is due to the grouping mechanism: if players are

⁷ The entire sample realizations (also for case 2) can be found in the Appendix.

⁸ I.e. a version of the game described above where agents only play one round.

⁹ Snapshots illustrating the evolution of the wealth distribution as a function of time can be found in the Appendix.

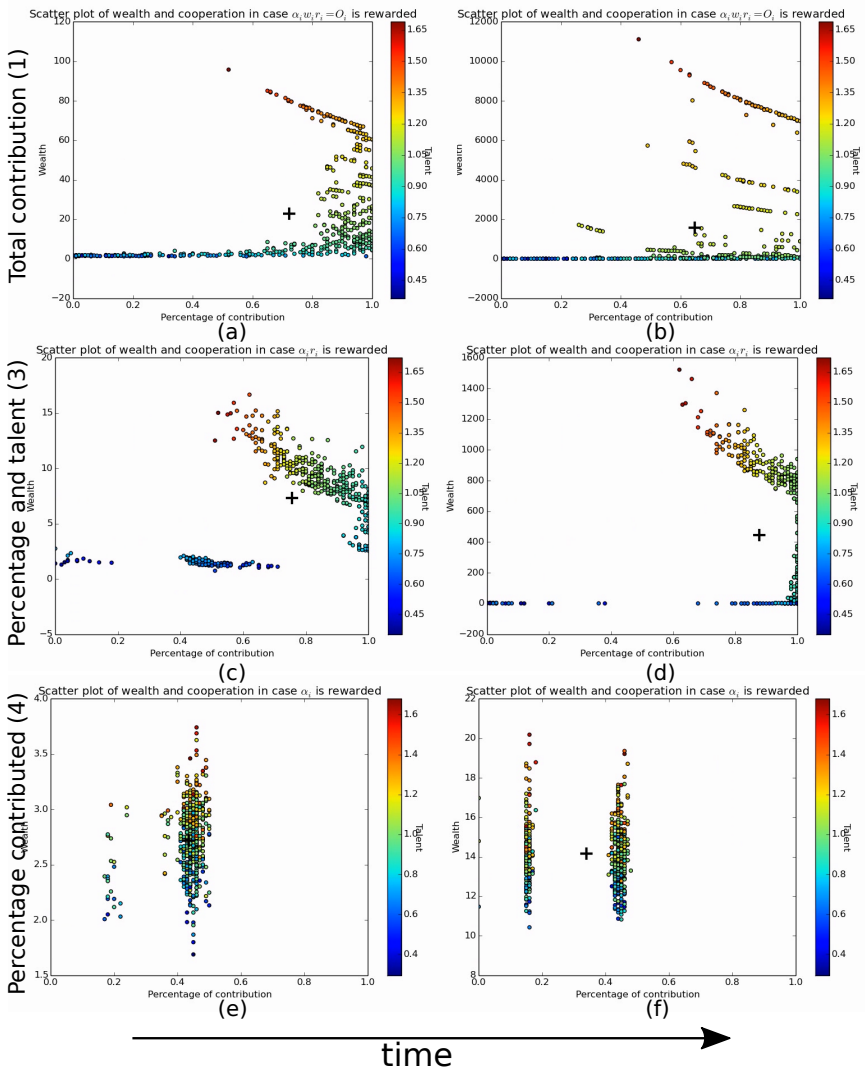


FIGURE 4.1: (Caption on the following page.)

FIGURE 4.1: Snapshots of sample realizations of the simulation described above for three different ranking criteria: players ranked based on their total contribution (case 1, first row), players ranked based on the percentage of their contribution taking into account their talent (case 3, second row), and players ranked solely based on the percentage of wealth they contributed (case 4, third row). Each scatter plot depicts the entire population, with each agent represented by a dot, at an earlier (left column) and later (right column) stage of the simulation. The abscissa of an agent indicates which percentage of his wealth he contributed to the common pool while his ordinate indicates his current wealth. The black cross marks the population average. The color of each agent indicates how talented he is, ranging from low talent (blue) to high talent (red). The simulation was obtained with the following set of parameters: $N = 500$, $S = 4$, $Q = 0.3$, $\sigma = 0.22$, $\lambda = 20$.

ranked based on more than just their chosen contribution level, it follows that highly endowed individuals (be it by wealth or talent) can percentage-wise contribute less and still be placed in a good group (meaning a group with a large common pool). For example, in the case of ranking 1, the wealthiest and most talented agent only has to contribute a percentage of his wealth so that his total contribution matches the maximum possible total contribution of the S th best player, to be guaranteed to be placed in the best group.

It is also possible to observe that early in the simulation, case 1 (fig. 4.1 (a)) and case 3 (fig. 4.1 (c)) are quite similar: many agents contribute a high percentage of their wealth to the common pool of the group. Agents with a low talent contribute significantly less than the population average and are also the ones with the lowest wealth.

Later in the simulation (fig. 4.1 (b) and (d)), the situation changes and the figure shows that the two ranking criteria lead to dramatically different outcomes: When players are ranked based on $\alpha_i \cdot t_i$ (case 3), the structure of the population later in the game (fig. 4.1 (d)) resembles the one earlier on (fig. 4.1 (b)): most agents contribute a higher percentage of their wealth to the common pool and wealth is distributed in a similar fashion as earlier in the simulation¹⁰. Instead, when players are ranked based on the total of their contribution (case 1), the situation later in the simulation (fig. 4.1 (b)) is quite different from the one in the beginning (fig. 4.1 (a)): less agents are contributing a high percentage of their wealth to the common pool and,

¹⁰ See footnote 9.

more importantly, the distribution of wealth becomes more unequal and it is possible to observe a clear separation in levels of wealth in the agents' population.

Indeed, as it was similarly observed in non repeated games [39] and chapter 2, the cause for the decline in contributions is the extreme inequality among agents: (relatively) poorer agents are disincentivized to commit a high percentage of their wealth to the common pool because it is in any case very difficult for them to match the total contributions of richer agents, even if they only contribute very little percentage-wise. Furthermore, the wealthiest among the agents also have less incentives to contribute to the common pool, because the bigger the difference in wealth, the lower is the benefit for wealthy agents to pool resources together with the less wealthy.

The differences in long term outcomes of the different ranking criteria are clearly visible in fig. 4.2. The figure displays the average of the population efficiency¹¹ (i.e. the ratio of the wealth realized in a given round over the maximum that could have been obtained, if every agent fully contributed to the common pool) over several simulations as a function of time for the four different ranking criteria: case 1 in blue, case 2 in green, case 3 in red and case 4 in cyan.

After an initial growth in efficiency (and hence in contributions) comparable for all cases, one observes that when agents are ranked solely based on their percentage-wise contribution α_i , the efficiency of the population quickly stabilizes around low values (cyan line in fig. 4.2). Instead, the ranking criteria that take into account the agents heterogeneity result into a much higher growth for the population efficiency (blue, green and red lines in fig. 4.2).

However, after a faster initial growth compared to case 3 (red line), in the cases where players are ranked taking into account the total amount of wealth contributed, the efficiency of the population starts to decline (blue and green lines in fig. 4.2). Indeed, as discussed above the high inequality of wealth in the population seems to result in diminished contribution levels over time, thus leading to a decline in the overall efficiency. Hence, on the long run, it appears that ranking agents taking into account the differences in their talent, but not in their wealth, leads to the best performance of the overall population.

Efficiency might not be the only relevant dimension when judging the performance of a system. Often, from a social planner point of view, one has two goals when considering the welfare properties of a system: to

¹¹ See the Methods section for a discussion on the efficiency measure.

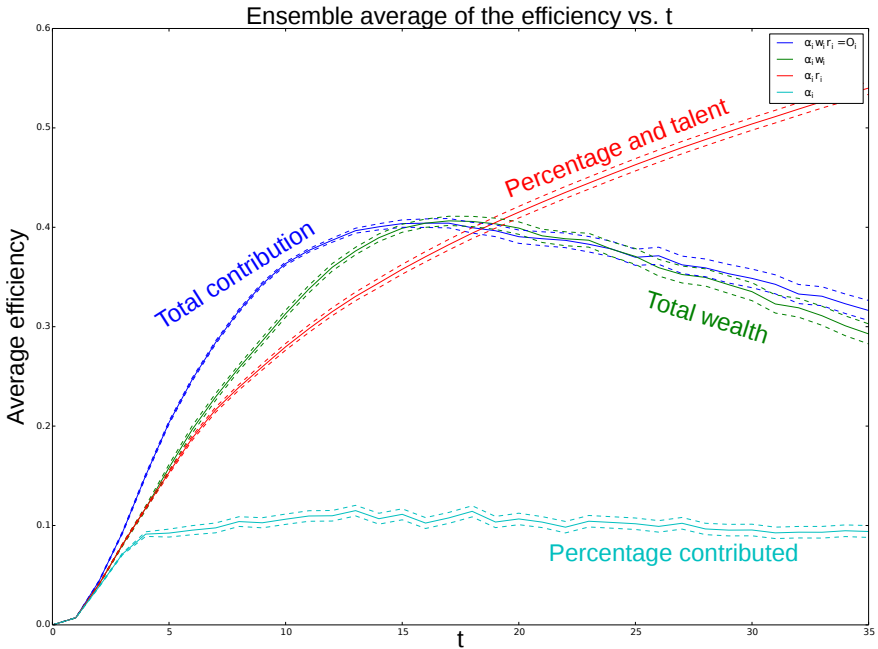


FIGURE 4.2: Average efficiency as a function of time for all the ranking criteria: case 1 in blue, case 2 in green, case 3 in red and case 4 in cyan. The results are obtained averaging over 2000 simulations and the dotted lines represent the 95% confidence interval for the ensemble average. After an initial growth in efficiency comparable for all cases, one observes that when agents are ranked solely based on their percentage-wise contribution α_i , the efficiency of the population quickly stabilizes around 10% (cyan line). Instead, the ranking criteria that take into account the agents heterogeneity result into a much higher growth for the population efficiency (blue, green and red lines). However, after a faster initial growth compared to case 3 (red line), in the cases where players are ranked taking into account the total amount of wealth contributed, the efficiency of the population starts to decline (blue and green lines). The results were obtained with the following set of parameters: $N = 500$, $S = 4$, $Q = 0.3$, $\sigma = 0.22$, $\lambda = 20$, $T = 350$.

maximize efficiency and to minimize inequality (see e.g. [114, 115, 117] for a discussion). However, as Okun [149] famously remarked: very often pursuing equality can reduce the overall efficiency of a system. It is thus of importance, when examining the different ranking criteria, to also look at how the wealth is distributed among the agents.

Figure 4.3 shows how the Gini coefficient¹² (a measure for wealth inequality in a population ranging from 0, meaning everyone has the same wealth, to 1, meaning that one agent owns the entire wealth of the population) evolves over time for the four different ranking criteria.

As expected, one can observe that ranking agents only based on the percentage of their wealth α_i contributed to the common pool (case 4), leads to low levels of wealth inequality in the population (cyan line in fig. 4.3). Barring the very beginning of the simulation, one also notices that assigning players to groups taking into account the total amount of wealth contributed to the common pool (cases 1 and 2, blue and green lines in fig. 4.3) leads to consistently higher inequality than ranking agents taking into account the differences in their talent but not in their wealth (case 3, red line in fig. 4.3). This is despite the fact that, on the long run, case 3 leads to higher amounts of wealth produced by the population.

Hence, from the point of view of a social planner who cares about equality as well as efficiency, choosing ranking criteria 1 and 2 would indeed be suboptimal¹³. I.e. one could say that ranking criteria 1 and 2 are “worse” than ranking criteria 3 and/or 4, in terms of the efficiency-equality trade-off. Of course, which ranking is “best” between 3 and 4, will depend on the social planner’s relative weight on efficiency and equality.

4.4 DISCUSSION

How to deal with “social dilemmas” has been a long-standing question in the social sciences. Contribution-based group formation promises to result in high levels of cooperative behaviour when applied to multiple-groups one-shot voluntary contribution games with equal agents. When played repeatedly, however, the outcome of the game might be quite different, especially when wealth accumulation is taken into account.

Using an agent based simulation, this paper investigates repeated VCGs with heterogeneous agents which accumulate wealth over time (thus in-

¹² See the Methods section for a discussion of the Gini coefficient.

¹³ More precisely: assuming a social planner assigning positive weights to efficiency and equality measure, one could say that cases 1 and 2 are Pareto suboptimal.

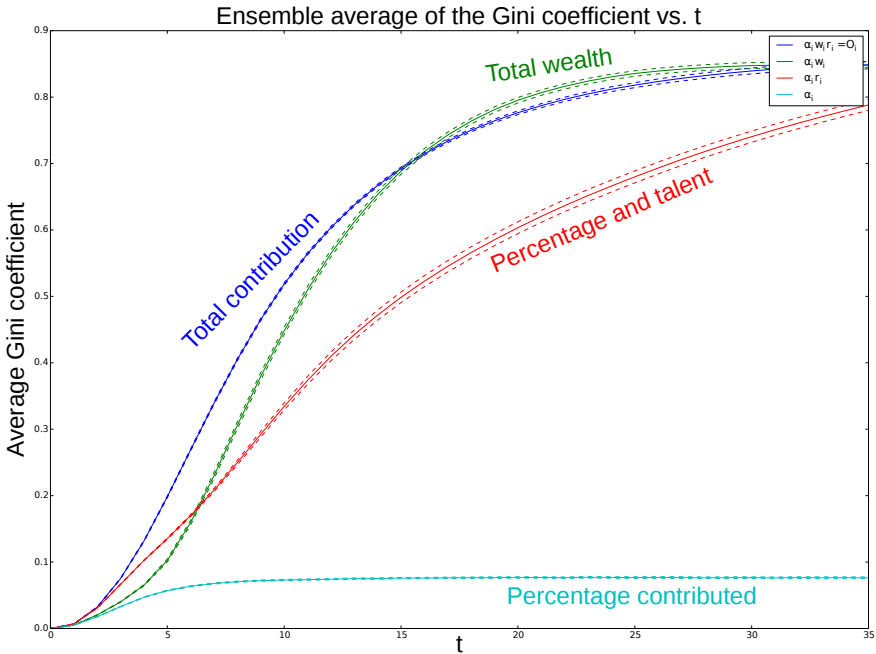


FIGURE 4.3: Average Gini coefficient as a function of time for all the ranking criteria: case 1 in blue, case 2 in green, case 3 in red and case 4 in cyan. As expected, ranking agents only based on the percentage of their wealth α_i contributed to the common pool (case 4), leads to low levels of wealth inequality in the population (cyan line). Barring the very beginning of the simulation, one also notices that assigning players to groups taking into account the total amount of wealth contributed to the common pool (cases 1 and 2, blue and green lines) leads to consistently higher inequality than ranking agents taking into account the differences in their talent but not in their wealth (case 3, red line). This is despite the fact that, on the long run, case 3 leads to higher amounts of wealth produced by the population. The results are obtained averaging over 2000 simulations and the dotted lines represent the 95% confidence interval for the ensemble average. The results were obtained with the following set of parameters: $N = 500$, $S = 4$, $Q = 0.3$, $\sigma = 0.22$, $\lambda = 20$, $T = 350$.

creasing in diversity). Whether, and at which level, cooperative behaviour emerges, depends on which criterion is used to assign people to groups. The matching criterion also starkly determines how the wealth produced during the game is distributed among the population. This paper studies several ranking criteria based on different dimensions and asks what are the properties of these mechanisms in terms of total production of wealth and its distribution. While in general it is impossible to determine a single best ranking system, some perform objectively better than others from the point of view of a social planner placing positive weights on the efficiency and equality of a population. In particular, it is found that ranking agents considering the total amount of wealth that they are able to contribute to the common pool, leads to a sub-optimal scenario in the long run.

On a broader level, the presented results show that it is not trivial to maintain in a more general setting the positive predictions obtained in the homogeneous one-shot assortative matching VCGs. An important component of the model considered here is which behavioural rules are followed by the agents. This paper assumes that agents update their strategies over the course of the game following a logit-response rule. Considering more realistic agents' behaviours (with the possibility of different agents following different behavioural rules) appears to be an exciting and important avenue for future work, in order to assess the robustness of the "good" equilibria of the game.

4.5 METHODS

Pseudo algorithm

To simulate the game described in the Model section , we used the following algorithm:

Algorithm 2: Pseudo algorithm for dynamical VCG with assortative matching

1 Initialization:

- Assign the same initial wealth to each player
- Assign talent to each player sampling it from a Normal distribution with mean 1 and standard deviation σ^a .
- Start the simulation in a fully defective state; i.e. set initial contributions α_i to 0 $\forall i^b$.
- Set time t to zero.

repeat

2 Update agents' contributions:

for $i = 1$ **to** N **do**

Generate random number X uniformly in $[0, 1)$.

if $X < p$ **then**

Do not update player i 's strategy.

else

Player i chooses a new contribution level α_j based on the logit probability distribution^c: $P_j = \frac{\exp(\lambda EU_j(\alpha_{-i}))}{\sum_k \exp(\lambda EU_k(\alpha_{-i}))}$, where λ is the rationality parameter and $EU_j(\alpha_{-i})$ is the expected payoff of strategy j given α_{-i} , the strategies played by all the other players in the previous round.

3 Group formation: Based on the contribution of each agent, groups are formed depending on the ranking criterion chosen, as described in the Ranking criteria subsection 4.2.1.

4 Materialize payoff: Each agent receives a payoff depending on the group to which he was assigned as described in equation 4.1.

5 Update wealth: Each agent adds the rounds's payoff to his portion of wealth that wasn't committed to the common pool. The time t is then increased by one.

until $t < T$;

^a Sigma is chosen small enough so that the VCG is still a social dilemma.

^b The interval $[0, 1]$ is discretized for computational purposes.

^c For a definition, see, e.g., [82].

In order to make sure that agents would converge to a Nash equilibrium¹⁴, inertia (i.e. updating the strategy only if a random number is above a certain threshold) was added when updating the player strategies (step 2 in Alg. 2)¹⁵. In this way, the algorithm is guaranteed to converge to one equilibrium, instead of oscillating around it.

The parameter λ is the so called *rationality parameter* and it determines what is the probability of an agent playing a strategy other than the one that would maximize his gains in the current round (assuming that no other player changes strategy from the previous round). A value of λ approaching infinity would result in an agent always playing the myopic optimal strategy, while for $\lambda \rightarrow 0$, an agent would choose a contribution level uniformly at random. A positive but finite value of λ allows agents to explore the strategy space while giving more weight to strategies that maximize their payoff. In this way, agents can converge to "good" equilibria and avoid to be stuck in "bad" ones; e.g. agents can converge from an equilibrium where no one contributes to one where some people contribute and are placed in good groups, thus earning a higher payoff. The results presented above hold qualitatively for a large set of values of the λ parameter. For very large values of λ ($\lambda \gg 100$), one can observe a qualitative difference¹⁶ in the outcome of the simulation when ranking according to percentage of wealth contributed and talent (case 3): Because of the extremely low exploration of the strategy space, highly talented agents do not manage to coordinate into concurrently contributing a high percentage of their wealth, thus failing to reap the benefit of being placed in highly profitable groups. Hence, after an initial increase in cooperative behaviour, the average percentage of contribution settles on the same level as in the case where agents are ranked only base on which percentage of their wealth they contribute (case 1), in striking contrast with what observed in figure 4.2.

4.5.1 *Efficiency and Gini coefficient*

4.5.1.1 *Efficiency measure*

The efficiency E of a population of N players at any given time step t is defined as the ratio of the sum of the payoffs of all players at time t over the

¹⁴ See e.g. [4, 76] for a definition of Nash equilibrium.

¹⁵ See e.g. [150] for a discussion on this topic.

¹⁶ Figures displaying average efficiency and Gini coefficient for $\lambda = 400$ can be found in the Appendix.

maximum sum of payoffs that could have been achieved in the same round, i.e. the sum of payoffs if every agent contributed his entire endowment to the common pool of his groups. Hence:

$$E = \frac{\sum_{i=1}^N \phi_i(\alpha_i)}{Q \sum_{i=1}^N w_i r_i} = \frac{\sum_{i=1}^N [w_i(1 - \alpha_i) + Q\alpha_i w_i r_i]}{Q \sum_{i=1}^N w_i r_i}$$

Therefore, efficiency is effectively a weighted measure of cooperation, with wealthier and/or more talented individuals having a bigger weight. A figure depicting the average cooperation as a function of time for all the different ranking criteria can be found in the Supplementary Material.

4.5.1.2 Gini coefficient

The Gini coefficient is a measure meant to represent the wealth distribution in a population of agents [151, 152]. A value of the Gini coefficient close to 0 indicates that wealth is distributed almost uniformly in the population while a value close to 1 means that almost the entire wealth of the population is owned by one individual¹⁷.

The following formula is used to compute the Gini coefficient G (see e.g. [rutherford1955income, 153] for the derivation and a discussion):

$$G = \frac{1}{N} \left(N + 1 - 2 \left(\frac{\sum_{i=1}^N (N + 1 - i)w_i}{\sum_{i=1}^N w_i} \right) \right) = \frac{2\sum_{i=1}^N iw_i}{N\sum_{i=1}^N w_i} - \frac{N + 1}{N}$$

with the N agents in the populations ordered from poorest to richest (i.e. so that $w_i \leq w_{i+1}$).

¹⁷ To put the numbers in perspective most OECD countries have a pre-taxes Gini coefficient around 0.5 and an after taxes coefficient below 0.35.

GROUPS AND SCORES: THE DECLINE OF COOPERATION

ABSTRACT

Cooperation amongst unrelated individuals in social-dilemma-type situations is a key topic in social and biological sciences. It has been shown that, without suitable mechanisms, high levels of cooperation/contributions in repeated public goods games are not stable in the long run. Reputation, as a driver of indirect reciprocity, is often proposed as a mechanism that leads to cooperation. A simple and prominent reputation dynamic functions through scoring: contributing behavior increases one's score, non-contributing reduces it. Indeed, many experiments have established that scoring can sustain cooperation in two-player prisoner's dilemmas and donation games. However, these prior studies focused on pairwise interactions, with no experiment studying reputation mechanisms in more general group interactions. In this chapter, we focus on groups and scores, proposing and testing several scoring rules that could apply to multi-player prisoners' dilemmas played in groups, which we test in a laboratory experiment. Results are unambiguously negative: we observe a steady decline of cooperation for every tested scoring mechanism. All scoring systems suffer from it in much the same way. We conclude that the positive results obtained by scoring in pairwise interactions do not apply to multi-player prisoner's dilemmas, and that alternative mechanisms are needed.

5.1 INTRODUCTION

Social dilemmas are situations where the optimal decision from the perspective of a self-interested individual conflicts with what is optimal for the group collectively. Examples include public goods [91] and common-pool resources situations [30], as modeled using game theory via, for example, prisoner's dilemmas (PD), voluntary contributions games [15, 21] or donation games [49]. The common feature of these interactions is that in the absence of a suitable mechanism [154, 155] and given insufficient foresight by the players [156, 157], the only stable outcome coincides with the socially undesirable one, i.e. absence of cooperation¹. The players fail to cooperate and, as a result, are all worse-off than in the collective optimum; a phenomenon often referred to as the "tragedy of the commons" [7, 8] or the "free riding dilemma" [158].

One of the most important mechanisms that successfully implements cooperation is "reciprocity" [159, 160]. Reciprocity is a behavior whereby people return benefits for benefits (and hostility with hostility) [51]. Thus, cooperation breeds cooperation and may lead to higher payoffs in the long run, if people resist the momentary benefits of defection (which, instead, breeds more defection and eventually leads to low payoffs). Commonly, one distinguishes between direct and indirect reciprocity. Direct reciprocity assumes that a player would cooperate with another person expecting he to do the same in return [145]; under indirect reciprocity, instead, a person does not expect the recipient of his help to reciprocate but he expects that someone else will [49]: "the recipients of an act of kindness are more likely to help in turn, even if the person who benefits from their generosity is somebody else" [161].

A principal driver of indirect reciprocity is reputation [162], therefore considered as a "universal currency" [60]: cooperating, or refusing to do so and choosing to defect, not only affects one's stage-game payoff but also one's reputation. When interacting again in the future, players will take each others' reputations into account, thus indirectly reciprocating players who have good a reputation (i.e. that have cooperated in the past). This creates incentives to cooperate beyond the momentary temptations of defection, provided the future benefits of cooperation are substantial. As a result, cooperation may emerge in the presence of suitable reputation mechanisms.

¹ In the remainder of this document, we will use cooperate as a common terminology for related terms like contribute, donate, exert effort, etc.

Indeed, reputation –via numerous implementations– has been shown to stabilize high levels of cooperative behavior in controlled experiments involving human subjects [163–165]. However, an important limitation of prior studies has been the focus on pairwise interactions, while in reality most social interactions unfold in groups [166] involving team production [167]. Producing in teams is particularly relevant in present society as interactions increasingly take place online, involving largely impersonal, crowd interactions.

Moving from pairwise interactions to group interactions substantially complicates matters in theory and in practice. In a group interaction, players might not be able to observe the actions undertaken by others individually, thus making it harder to track and update other players' reputations. Other than in a two-person interaction, one can often not infer the others' individual actions from the aggregated outcome. For instance, when playing a public good game, information regarding individual behaviors may not be available, and the only available information may concern the group as a whole.

This raises the following question: How do reputation mechanisms fare in group interactions? More specifically, as a first step towards addressing the question more generally, we shall here investigate one of the best-known and simplest mechanisms for reputation called "scoring". Our analysis of "group scoring" extends the concept of "image scoring" [62, 66], as has been studied widely in pairwise interactions. Under image scoring [49, 61], each player has a score (starting at 0) as a proxy for his reputation. Whenever a player has the opportunity to cooperate with someone else, his score is updated: if he cooperates his score is increased by one, if not it is decreased by one. Thus a player's reputation is continuously reassessed based on the past (in the simplest case, based on the previous decision). A seminal theory result [61] is that the strategy to cooperate with anybody with a non-negative image score is evolutionary stable. Crucially, by refusing to cooperate with someone with a low image score a player is decreasing his own score, thus reducing his own probability of receiving cooperation in the future. Hence, not cooperating with a player with a low image score can be interpreted as a form of punishment. Indeed, in practice, numerous behavioral experiments show that image scoring helps stabilize cooperative behavior in two-player PDs and donation games [66–68, 168].

As we extend scoring mechanisms to group interactions more generally, and to multi-player PDs in particular, we increase the degree of freedom

regarding the scoring rules that may apply. Real-world group interactions vary with respect to the information that is available, and typically individuals do not observe all actions undertaken by all other individuals, especially in large groups. The relevant scoring mechanism that applies to a specific group interaction therefore depends on how much information is available to players and how much information each reputation rule requires, as processing the available information correctly may become difficult in larger interactions. Indeed, a conjecture [60] for why image scoring is favored over other reputation dynamics is that (relatively) little information is required to implement it under full feedback [70]. As such, with limited [69] or partially erroneous feedback [169], sufficiently accurate information is key for mechanism success.

When interacting in groups, information becomes coarser and a single subject may thus find it harder to reap the benefits of “reputation-building”, and cooperation may therefore unravel. Recent theory has extended “scoring” methods to group interactions [71]. The baseline establishes a positive cooperation result for the case of image scoring in group interactions². Furthermore, when only information regarding group performance –but not regarding individual players– is available, “group scores” replace image scores: each player’s group score summarizes the aggregate cooperativeness of the groups to which he belonged in the past, without any additional information regarding what players did individually. In this case, theory predicts that cooperation cannot be sustained.

In this chapter, we provide the first test of this theory in a group setting considering various informational contexts. Hence, as a first step toward addressing this question more generally, we investigate whether different scoring mechanisms can sustain cooperation in experimental multi-player PDs. In particular, we consider a simple and widely used implementation for scoring mechanisms based on ‘Markovian’ scores, that is, scores that depend only on players’ actions from the previous period (“memory 1”). The basic model we consider is an individual-level binary³ Markovian ‘image score’, as investigated theoretically in numerous prior studies (e.g. [49, 62, 65, 170–173]). For such scores, theory predicts that high levels of cooperation can stabilize, and there exists experimental evidence confirming this in the context of pairwise interaction [69, 174]. In fact, concerning the role of memory, existing experimental evidence [174] suggests that Markovian memory already leads to high levels of cooperation and that longer

² I.e. when full information regarding individual decisions is provided.

³ Meaning that the score of a player can only have two values, e.g. 0 and 1.

memory increases cooperation further. The goal of the present chapter is to investigate whether, for the case of the Markovian baseline, the positive results that were obtained for pairwise interactions carry over to group interactions.

For this, we conducted an extensive laboratory experiment. The baseline is to test image scoring. In addition, we test alternative scoring rules that could apply to group interactions including one where players score each other endogenously through votes. The proposed rules differ with respect to how much information regarding past behavior of their group-mates is required, ranging from no feedback to full feedback.

The experimental results concerning cooperation are negative: for every scoring mechanism we observe a steady decline in cooperative behavior. The decay of cooperation is the same under every mechanism and comparable even with the case when no scoring mechanism is implemented at all. We conclude that positive results regarding cooperation deriving from scoring, as were repeatedly observed in two-player interactions, do not generalize to group interactions. Our results confirm the negative theoretical prediction with respect to coarse group scoring but falsify the positive prediction regarding image scoring in groups.

The rest of this chapter is structured as follows. Next, we present the experimental procedure, followed by our results. A Methods section contains additional details concerning experimental design and statistical analyses.

RESULTS

Before presenting results, we briefly discuss the structure of the experiment and introduce the different scoring mechanisms that were tested. For further detail concerning the experimental design, we refer the reader to the Methods section and appendix C.

Experimental procedure

Our experiment involved 192 subjects playing several, repeated multi-player PDs, resulting in a total of 11,520 on whether to cooperate or not. The experiment had 12 sessions involving 16 subjects each; each session consisted of three different treatments, each played for blocks of 20 rounds (“phases”). In each treatment, subjects were faced with a different scoring mechanism and treatments differed according to which and in which order the following mechanisms were implemented:

Scoring mechanisms

Treatment	Feedback provided
No scoring	No feedback about other players' actions
Image scoring	Feedback about individual actions of others
Group scoring	Feedback about average behavior in the group
Self scoring	Endogenous feedback
Image self scoring	Same as Image scoring (control for Self and Image scoring)

TABLE 5.1: **Summary of scoring mechanisms:** The above table summarizes how much information about other players' actions in the previous round was provided to the players. Regardless of the treatment, all subjects were given feedback regarding the profit made during the round (and hence on the number of contributors in their group).

Scoring mechanisms range between image scoring, providing full feedback about other player's actions, and no scoring, providing no feedback at all:

- **No scoring:** Subjects receive no information at all regarding the past actions of the other players, and therefore it is the treatment with the lowest informational content. *Expectation:* In this implementation of a repeated multi-player PD we expect a decay of cooperation resulting in low contribution levels, as shown by numerous previous experiments [66, 68, 70, 175] mainly conducted in voluntary contribution games settings.
- **Image scoring:** This is the treatment with the highest informational content of all, equivalent to the case with a binary image score in two-players interactions. Players are told whether their past and future group-mates cooperated in the previous round. *Expectation:* Based on previous experiments on donation games [175] and on theoretical results [71], one could expect a stable high level of cooperation.
- **Group scoring:** Scoring proceeds as in image scoring, except that all group members receive the same score based on the number of cooperators in their group. Subjects are given no direct information about individual decisions. *Expectation:* Recent theoretical work [71] suggests that a low level of cooperative behavior is to be expected.

- **Self scoring:** Players directly assign the score to their fellow players based on feedback regarding own payoffs and aggregate contributions in their group. This treatment might contain more or less information than group scoring depending on whether players are truthful when assigning the scores. *Expectation:* In this case the only Nash equilibrium is for nobody to contribute, independently of the assigned ratings.
- **Image self scoring:** This is a control treatment for self and image scoring, where scores are exogenously assigned as if all the players were truthful in the Self scoring treatment. The resulting informational content is, in principle, equivalent to Image scoring, but provided in a slightly more complicated format.

Every round, subjects were randomly reshuffled and rematched into groups of size 4 and provided with scores feedback, in particular of their group-mates, calculated using the current scoring rule. After deciding whether to cooperate or not, subjects received their personal individual payoff feedback (thus knowing how many people cooperated in their group) and were assigned updated scores.

It is important to note that, by virtue of our design, the score of a subject only reflected his last action, and that scores did not carry over multiple rounds of the game. Our focus is on situations where mechanisms are introduced or where a new mechanism replaces an old one. Hence, subjects in our experiments always initially played a treatment where no feedback about others' actions or scores was given ("Initial phase"). After that, two different scoring mechanisms were played in succession ("Scoring phase 1" and "Scoring phase 2").

Experimental results

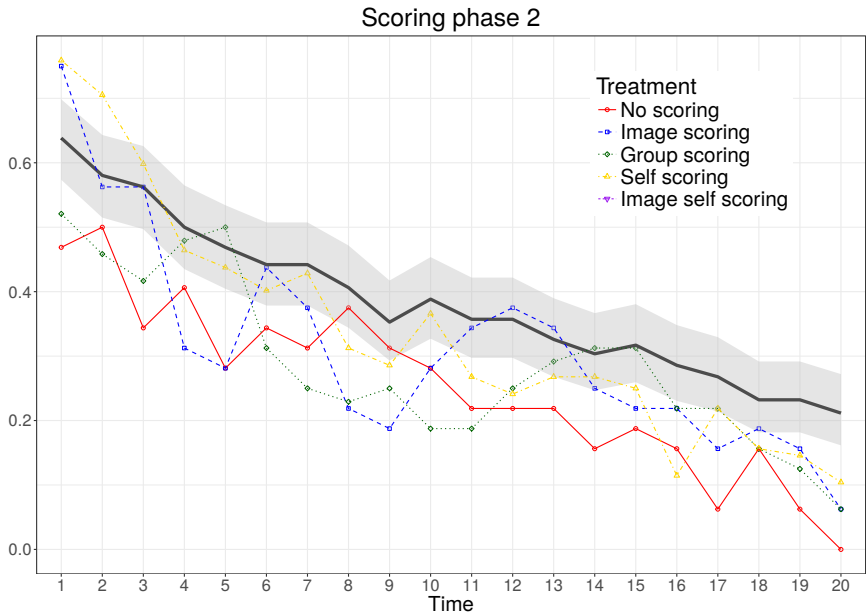
In fig. 5.1 we show the percentage of cooperators as a function of time for all the different treatments⁴. For all treatments, we observe a steady decline in cooperation; the decay occurs in much the same way, independent of the order in which the different treatments were played⁵.

⁴ More detailed plots are available in Appendix C.

⁵ This decay is in line with similar patterns known from multi-player public goods games (see e.g. [15, 19, 110, 176, 177]).



a



b

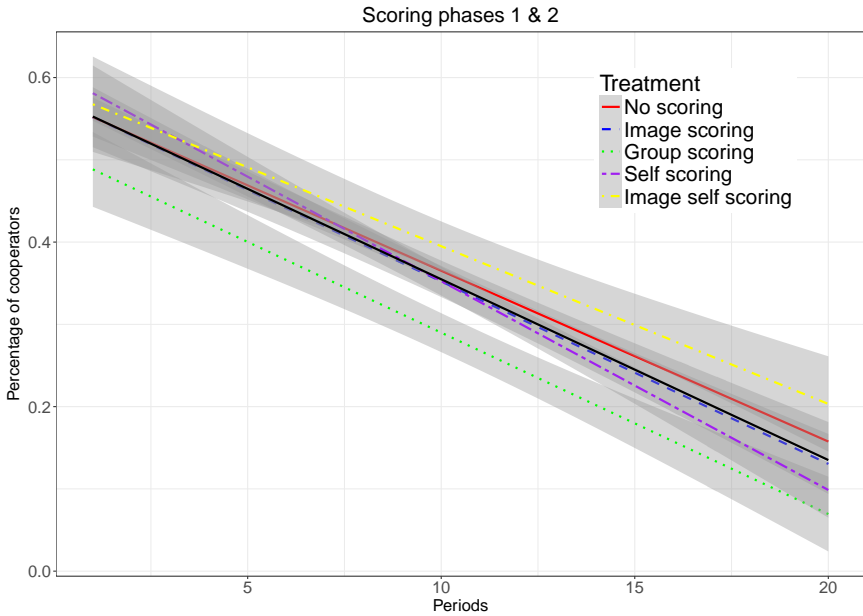
FIGURE 5.1: (Caption on the following page.)

FIGURE 5.1: Percentage of cooperation as a function of time for all the treatments: The figures on the top and on the bottom show the contribution levels observed during the first and second scoring phases of the experiment, respectively. The black line in the background shows the average cooperation observed in the initial phase. Since first treatment subjects (i.e. in the initial phase) always played the treatment with no scoring mechanism, it can be treated as a baseline. The grey area represent the binomial proportion confidence interval [178]. The figures show a steady decline in average cooperation. The decline happens in much the same way for all treatments, and independent of the order in which the different treatments were played.

Even though there are significant statistical differences between some of the observed downward trends (e.g. image scoring is significantly different from no scoring, see table 5.2), the main difference in treatments can be reduced to a slight offset in the initial percentage of contributors. Figure 5.2 illustrates that the estimated (linear) decay of cooperation over time occurs at the same speed. Indeed, all the slopes are within the error range of each other. The only noticeable difference regards the intercept, that is, the initial contributions (see fig. 5.2 (b)).

	No scoring	Image scoring	Group scoring	Self scoring	Image self scoring
No scoring	n/a	0.063	0.651	0.043	0.029
Image scoring	0.063	n/a	0.158	0.870	0.741
Group scoring	0.651	0.158	n/a	0.115	0.082
Self scoring	0.043	0.870	0.115	n/a	0.867
Image self scoring	0.029	0.741	0.082	0.867	n/a

TABLE 5.2: **Pairwise Mann–Whitney–Wilcoxon Rank Sum tests.** The table shows the p-values obtained from the Mann–Whitney–Wilcoxon Rank Sum test for each pair of treatments. The test was performed only on first-period decisions and excluding the initial phase of every session. Red p-values indicate a statistically significant difference (with dark red when $p < 0.05$ and light red when $p < 0.1$) while a black p-value indicates no significant difference.



a

Treatment	Estimated slope
<i>All data</i>	-0.022 ± 0.001
<i>No scoring</i>	-0.020 ± 0.001
<i>Image scoring</i>	-0.022 ± 0.002
<i>Group scoring</i>	-0.022 ± 0.002
<i>Self scoring</i>	-0.025 ± 0.002
<i>Image self scoring</i>	-0.019 ± 0.003

b

FIGURE 5.2: **Estimated decays of cooperative behavior.** In figure (a), each colored line illustrates the fitted linear function of a treatment. The grey areas depict the 95% confidence interval. The black line depicts the estimated decay for the entire data set. Table (b) lists the values obtained for the various slope estimators. There is a difference in some of the intercepts of the different lines, but all treatments decline with (statistically) similar slopes.

For more details on the statistical analysis, we refer the reader to the Methods section.

The above results indicate that the scoring mechanisms considered here, even ones which were shown to stabilize high level of cooperation in two-players games (i.e. image scoring), fail to achieve positive results in multi-player interactions. The most plausible explanation is that it is harder to isolate the “bad apples” in a group interaction, resulting in a deterioration of the quality of scores, as perceived by subjects. This kind of imprecision destabilizes cooperation: to keep stable levels of cooperation, players should –on average– cooperate with a frequency at least as high as the observed number of players with a high score in their group, thus maintaining a stable percentage of cooperators in the population. Instead, we observe that, while, *ceteris paribus*, players do cooperate more with an increased observed score in their group, they do so with a (downward) bias, especially for high sums of scores in the group⁶. Figure 5.3 illustrates the case of image and group scoring⁷: in the picture we can see that players cooperate less than 80% (on average) of what they should cooperate in order to obtain stable cooperation. This behavior is also confirmed by an analysis of individual decision making: subjects positively react to observed high scores in their group, but they do not “reciprocate” enough for cooperation to be stable. A formal model and analysis of the players’ decision making is presented in the Supplementary Material.

Further contributing to the steady decline of cooperation is the fact that when a high-score player decides not to cooperate because of the presence of low-score subjects in his group, this reduces the score of all his group-mates, not just of the low-score individuals. This results in a steady decay of players with good reputation and cooperative behavior in the population, and consequentially to a downward spiral of contributions, as observed by Fischbacher et al. [47] in a study on (imperfect) conditional cooperation in a public goods experiment.

6 It is important to note that this effect relies on the players being able to observe the scores in their group. If this is not the case, like in the “No scoring” treatment, no such effect is observed. See the Supplementary Material for more information.

7 See Supplementary Material for the other cases.

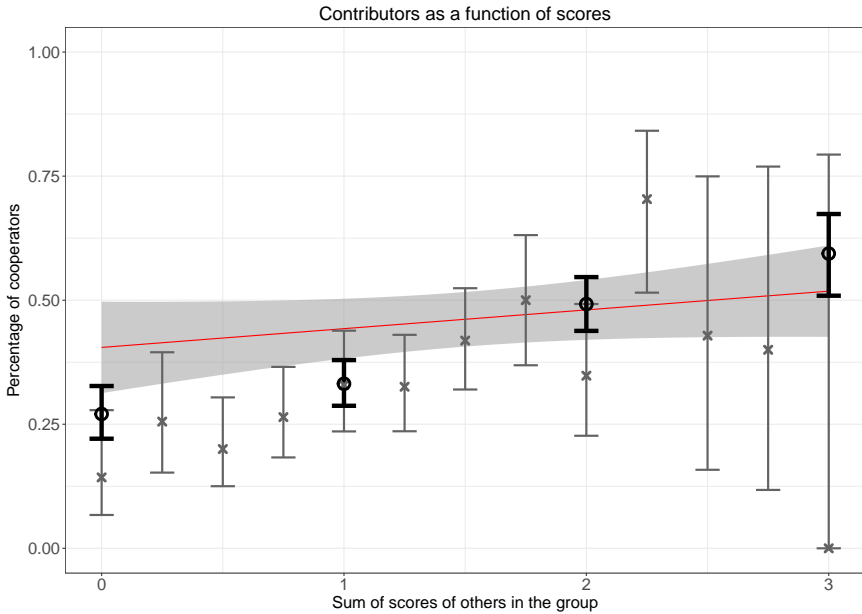


FIGURE 5.3: **Percentage of cooperators as a function of the observed score in their group.** The figure shows how many players contributed to the common pool of their group as a function of the sum of the scores in the group. Black and grey illustrate the image scoring and group scoring cases, respectively. The error bars indicate the 95% binomial proportion confidence interval computed using Wilson score intervals [178]. The red line depicts the percentage of cooperations in the no scoring treatment as a function of what *would have been* the observed score (see Supplementary Material for more details); the gray area represent the 95% confidence interval. The figure shows that, even though players cooperate significantly more when the observed sum of scores increases, they do so with a downward bias, as compared with the identity line, especially for high score values; decisions below the identity line will result in a steady reduction of “good players” in the population, thus lowering the average score and resulting in a spiraling down of cooperative behavior.

DISCUSSION

Scoring methods in general, and image scoring in particular, are simple implementations of reputation mechanisms. They stabilize cooperative be-

havior in various standard, two-players social dilemma situations, such as in prisoner dilemmas or donation games. Image scoring requires reliable feedback regarding individual-level behavior. The purpose of this study is to extend such mechanisms to group interactions, in particular to multi-players prisoner dilemmas with or without full individual feedback. We propose several scoring rules that could apply in this setting, depending on informational context, and test them in a laboratory experiment. Furthermore, we investigate how an endogenized scoring mechanism could be implemented. The results are unambiguously negative: independent of information, feedback and scoring mechanism, cooperation decays. This includes mechanisms that were previously shown to stabilize cooperation in corresponding two-player cases. A plausible explanation is that individuals cannot be isolated; i.e. defectors cannot be individually punished, and cooperators cannot be individually rewarded. This results in a reaction to the average group score that is increasingly biased toward defection, therefore leading to a steady decrease of high-reputation players in the population that in turn begets lower levels of cooperative behavior.

On a broader level, our results show that there is still much that we do not know about reputation dynamics. Even though indirect reciprocity is considered one of the main mechanisms through which cooperation can be sustained, there have been very few studies on interactions in group setting. Understanding such settings has become particularly relevant in recent years because, due to the increasing digitalization of our world, more and more interactions take place online where people frequently communicate via crowd platforms and where often explicit reputation tallying is provided as a method to build trust. Due to the increasing decentralization of interactions, partial or total anonymity of the actors involved can be the norm and reputation is often built on a peer-to-peer basis with members of communities rating each other. For example, a project may involve several groups of individuals, and information on individual level contributions could be imperfectly filtered via several community-layers before reaching the players. With this work we set up to investigate some of these issues.

A key conclusion is that many positive results on cooperation, as have been observed in pairwise interactions, may not hold anymore when groups are concerned.

There are numerous avenues for future work and many open questions; for example: are dynamics of play in multi-player games fundamentally different from two-player games (as it might be the case for direct reciprocity, see e.g. [179–181])? And if so, could one exploit this to devise a

scoring mechanism that is able to sustain higher levels of cooperation? How does group size matter? Could the combination of multiple mechanisms, such as scoring and punishment, lead to higher cooperation? Could the deterioration of the quality of scores be compensated by cumulating the scores over multiple rounds, letting players “build” their reputation? Future work should address such issues and many others, as group structures are an important, ubiquitous aspect of human society.

METHODS

The experiment

The experiment was conducted as an experiment on interactive decision-making at the ETH Decision Science Laboratory (DeSciL) in Zurich using the z-Tree [182] software. We ran 12 sessions with 16 participants in each session, for a total of 192 participants. Participants were recruited from the joint subject pool of ETH Zurich and University of Zurich using the hroot [183] software and mainly consisted of university students. All procedures adhered to DeSciL’s Operational Rules⁸; additional ethics approval was waived following standard DeSciL protocol for members of the laboratory’s Review Board. In no way at all does the experiment violate the ethical principles of the Declaration of Helsinki, and subjects were properly incentivized by converting their earnings in real currency with full transparency (i.e. no deception). Each session in the laboratory lasted roughly one hour during which the players played 3 treatments for 20 rounds each. On average, subjects earned 33 CHF (roughly 33 USD at the time), with a range of 25 to 40 CHF, including a 5 CHF show-up fee.

First, subjects always played the treatment were no information regarding past behavior was provided. After that, subjects played two of the other scoring treatments. The table below details the treatments’ combinations.

⁸ These implement the standards of behavioral economics including no deception, compatible incentives and payment, minimal earnings, rights to terminate experiments at any time, data anonymity, and confidentiality

	Treatments' combinations		
	Initial Phase (Round 1-20)	→ Scoring Phase 1 (Round 21-40)	→ Scoring Phase 2 (Round 41-60)
Control treat.	<i>No scoring → No scoring → No scoring</i>		
Treat. Com. 1	<i>No scoring → Image scoring → Group scoring</i>		
Treat. Com. 2	<i>No scoring → Group scoring → Image scoring</i>		
Treat. Com. 3	<i>No scoring → Image scoring → Self scoring</i>		
Treat. Com. 4	<i>No scoring → Image self scoring → Self scoring</i>		
Treat. Com. 5	<i>No scoring → Self scoring → Image scoring</i>		

TABLE 5.3: **Combinations of treatments played during the experiment:** Each row details one of the 6 treatment combinations in the experiment. Each combination was played twice (in two different experimental sessions).

At the beginning of the session and before each treatment, subjects were given written instructions⁹ explaining what the experiment was about and the game that they were about to play, scoring mechanism included. Before the first treatment, subjects were given some minutes to familiarize with the game with a small training. Before the Truthful self scoring treatment, because of the complexity of the scoring mechanism, subjects also had some minutes to understand how the scoring worked using a score simulator. Screenshots displaying the different treatments can be found in the Supplementary Material.

As customary, subjects were incentivized by converting their earnings in real currency. Subjects on average earned 33 CHF (roughly 33 USD), including 5 CHF of show-up fee. Earnings ranged from 25 to 40 CHF.

In the following we define the game that the subjects played in the experiment and the scoring mechanisms that were used.

N-players prisoner's dilemma

The subject played the following game whose aspects were all common knowledge:

⁹ Available in the Supplementary Material.

1. At the beginning of each round (for 20 rounds), N subjects ($N = |n|$ where $n \equiv \{1, \dots, 16\}$) are randomly assigned to four groups of fixed size four.
2. Every subject decides whether to contribute his endowment to the common pool (i.e. whether to cooperate). For player $i \in n$, let $c_i = 0$ and $c_i = 1$ denote whether player i cooperated or not, respectively. Starting from the second round of play, players are shown the scores assigned to all players in the previous round. Furthermore, players learn the score of their group-mates in the current round.
3. Subjects receive individual payoff ϕ according to $\phi_i = (1 - c_i) + \frac{1}{2} \sum_{j=1}^4 c_j$.
4. *Scoring*: a score is assigned to each player based on his contribution and depending on the treatment. The score is visible to the other subjects in the following round and it replaces the score from the previous round.

Regardless of the treatment, all subjects were shown the profit that they made during the round and during the entire session; thus each subject was told how many people cooperated in his group in the previous round.

The scoring mechanisms

Depending on treatment, a different score was assigned to each subject. The score was not cumulative over rounds and, every round, subjects were only shown the scores (if any) as were assigned in the previous round. The scoring mechanisms were designed so that the score ranged between 0 and 1 for all treatments.

- **No scoring:** No score was assigned to players during this treatment.
- **Image scoring:** Subjects were assigned a score of 1 if they cooperated in the previous round and 0 if not.
- **Group scoring:** Subjects were assigned a score proportional to how many people in their group contributed to the common pool. The score equaled the number of cooperators in their group divided by the group size (4); thus subjects in the same group all received the same score. More precisely, the score s_i of player i in group G_i equals as $s_i = \frac{1}{4} \cdot \sum_{j \in G_i} c_j$. In principle, the higher subject i 's score, the higher

is the probability that i invested in the group account. If the resulting score is 1 or 0, the group score faultlessly indicates whether a subject cooperated or not, respectively.

- **Self scoring:** Each subject was asked to rate his/her group awarding a number of stars ranging from 0 to 3. The score of each subject was computed as the sum of all the stars awarded to the group by his group-mates (excluding his own rating) divided by 9 (i.e. the maximum number of stars that a player could be assigned). Therefore, indicating with $\star_j \in \{0, 1, 2, 3\}$ the score assigned by player j in group G_i to his group, the score s_i of player i in group G_i was computed as $s_i = \frac{1}{9} \cdot \sum_{j \in G_i, j \neq i} \star_j$. Hence, the score of each subject ranges between 0 (all his group-mates awarded 0 stars to the group) and 1 (all his group-mates awarded 3 stars to the group).
- **Image self scoring:** The score was assigned as in the self scoring treatment but exogenously. This means that each subject was considered as having awarded a number of stars to his group equal to the number of cooperators (excluding himself) observed in his group. More precisely, for a group G_i we denote with $u_i \in \{0, 1, 2, 3\}$ the sum of the players cooperating in G_i as observed by player i ; i.e. $u_i \equiv \sum_{j \in G_i, j \neq i} c_j$. The score s_i of player i in group G_i was then computed as $s_i = \frac{1}{9} \cdot \sum_{j \in G_i, j \neq i} u_j$. Hence, each score ranges between 0 (all players in that group defected) to 1 (each player in that group cooperated).

Statistical Analysis

To determine if treatments significantly differ from one another, we used the Mann-Whitney-Wilcoxon rank sum test [184, 185]. Due to (possible) autocorrelations between same-session decisions, we restricted our analysis to only decisions in the first period. Furthermore, we exclude from the analysis decisions taken during the initial phase. Let $x_i^q \in \{0, 1\}$ denote the decision that player i took during the first period of treatment q . We obtain $\bar{x}^q \equiv \{x_1^q, \dots, x_m^q\}$ where m is the number of players that played treatment q (excluding the initial phase). We perform a rank sum test for each pair of treatments: the p-value obtained from the test is a measure of how likely it is that \bar{x}^i and \bar{x}^j are drawn from the same distribution with the same mode. Table 5.2 shows the p-values for each pair of treatments in

the first and second scoring phase of the experiment. A value depicted in red indicates that the two treatments significantly differ from each other.

To obtain fig. 5.2 we performed a linear regression of the contributions to the public good as a function of time for each treatment individually and for all of them combined. An alternative analysis, using a random resampling permutation test, is available in the Appendix. In Appendix C, we also provide a model for the decision making of the individual player and fit it to our data. The obtained results are compatible with the ones presented in this chapter.

CONCLUSION

You can't have your cake and eat it, too, is a good candidate for the fundamental theorem of economic analysis. We can't have our cake of market efficiency and share it equally.

— A. M. Okun

Understanding how cooperation is organised, distributed and maintained, among humans and in other situations, is one of the oldest and most interesting questions in game theory. The dilemma of the emergence of cooperative (or altruistic) behaviour can be summarized by the free-riders problem: The free-riders problem arises from the fact that, while an entire population benefits from the presence of a public good produced at some cost by cooperative individuals, free-riders (defectors) can still benefit from the public good without producing any of it. Exploiting the cooperative agents, the defectors save the cost of producing the public good, obtaining an advantage over cooperators. As a consequence, on the long run the players often fail to cooperate and, as a result, are all worse off than in the collective optimum; a phenomenon often referred to as the “tragedy of the commons”. Yet, in the real world, we often observe cooperative-like conducts in a variety of cases, ranging from human behaviour to bacterial traits. In the past years, much literature has been devoted to understanding why is cooperative behaviour sustained in situations where models predict it should not, and how could this behaviour be incentivized in situations where it is not.

In this dissertation, I focus on Public Goods games, and in particular in understanding under which conditions can the public goods be supplied through voluntary contributions when players interact in groups. I not only focus on how, through incentive mechanisms or exploiting behavioural regularities, cooperation can emerge, but on what are the implications of the needed mechanisms in terms of the total welfare of the society. In particular, I study Voluntary Contribution Games and focus on mechanisms that have been shown to lead to high cooperation without needing allow pay-off manipulations, such as transfers or subtractions of wealth. The aim is to assess, using a mixture of analytical, computational and experimental

tools, how robust are these positive predictions (in terms of cooperation) when more general models, thus closer to real-world social dilemmas, are taken into account.

I do this by first looking at the family of “grouping” mechanisms: these mechanisms have been shown to result in almost Pareto-efficient equilibria by ranking players based on their contributions and then ranking them accordingly. I first test what is the effect on the equilibria predictions of the game if different action spaces and a wider range of public goods provision efficacies are considered. I find that the equilibria do not depend much on the exact nature of the available action space but that they are very much dependent on “how good” the public-good provision efficacy is.

A central focus of this dissertation is to tackle a crucial limitation of previous scholarship: accounting for diversity among players. Using a mixture of analytical and computational tools, I show that the consequences of including heterogeneous agents in the model, while of course depending on the exact structure of the game, are often quite detrimental: Indeed, unless the action space of the game is very coarse, all the highly efficient equilibria realized in the case of homogeneous agents cease to exist; instead, the game reverts back to either the fully non contributive equilibrium, or to new, complex, highly inefficient mixed strategies.

I also address the case of imperfect grouping mechanisms by adding noise to the observed contributions and show that, if the noise is not too high, nearly efficient equilibria continue to exist whenever they exist in case of perfect matching (albeit for a smaller parameters’ set).

Building on the above results, I investigate how the equilibria of the games discussed above change when moving from the one-shot case to repeated interactions. I show that, due to wealth accumulation, the heterogeneity among players increases over time, thus leading to different equilibria than in the one-shot case. Having introduced agents’ diversity, different criteria based on which one could assign agents to groups are also automatically introduced. Interestingly, it is possible to show that the average cooperation among agents dramatically depends on which criterion it is used to rank the agents.

Furthermore, for both the one-shot and repeated games, I examine the welfare effect of the different mechanisms discussed above. Interpreting noise and ranking criteria as policy tools, I determine which choices would maximize the welfare in the population, from the point of view of a social

planner whose preferences place positive weight on maximal efficiency and minimal inequality among the population.

Lastly, I shift the focus on "scoring", a behavioural mechanism also known to achieve high level of cooperation in two-players games with repeated interactions, by implementing a reputation dynamic as a driver of indirect reciprocity. I extend scoring mechanisms to group interactions, considering the different degrees of available information that naturally emerge due to the grouping structure (e.g. players might only be able to observe the average performance of a group). I also investigate whether it is possible to decentralize the scoring mechanism by allowing players to themselves rate their fellow group members. Using a behavioural laboratory experiment, I test several scoring rules which apply to group interactions (differing with respect to how much information regarding past behaviour of other players was provided) and find unambiguously negative results: All the proposed scoring rule fail to sustain cooperation on a longer time-scale, even image scoring, i.e. the one that was experimentally shown to stabilize high level of cooperative behaviour in two players interactions.

This thesis only took some initial steps in the direction of giving answers to the fundamental questions concerning the existence and robustness of efficient equilibria of VCGs played in groups, and there is much that still needs to be addressed:

An important assumption of the models discussed above regarded which behavioural rules are followed by the agents. An exiting avenue for future work, both theoretical and experimental, is to consider more realistic agents' behaviours, in particular taking into account the effect of players following different behavioural rules while interacting with each other. A further possible research direction is to consider the effects of a better players' "memory": computationally, by implementing more sophisticated strategy selection dynamics than myopic best response, and experimentally, by testing for scoring mechanisms that allow players to "build" their reputation over multiple rounds. Finally, the combination of multiple mechanisms which individually are not able to sustain cooperation in groups but that might be able to do it jointly, should be investigated.

More broadly, all the mechanisms that were considered in this thesis try to incentivize contributions to a public good by implementing mechanisms that reward a "good" behaviour, be it via contribution-based grouping or reputation, and require no payoff transfers between players. However, the results presented in this dissertation show that there is still much that

we do not know about Voluntary Contribution Games (and Public Goods Games in general) when played in a group setting. Understanding such settings is of fundamental interest due to the fact that most real-world interactions take place in groups, thus making it crucial to extend our models to these situations in order to better apply whatever insight is gained to address real social dilemmas.

APPENDIX: ASSORTATIVE MATCHING WITH INEQUALITY IN VOLUNTARY CONTRIBUTION GAMES

A.1 EFFICIENCY LOSS AS A FUNCTION OF THE ENDOWMENT DISTRIBUTION

The following is an heatmap displaying the efficiency loss for different widths of the endowment distribution.

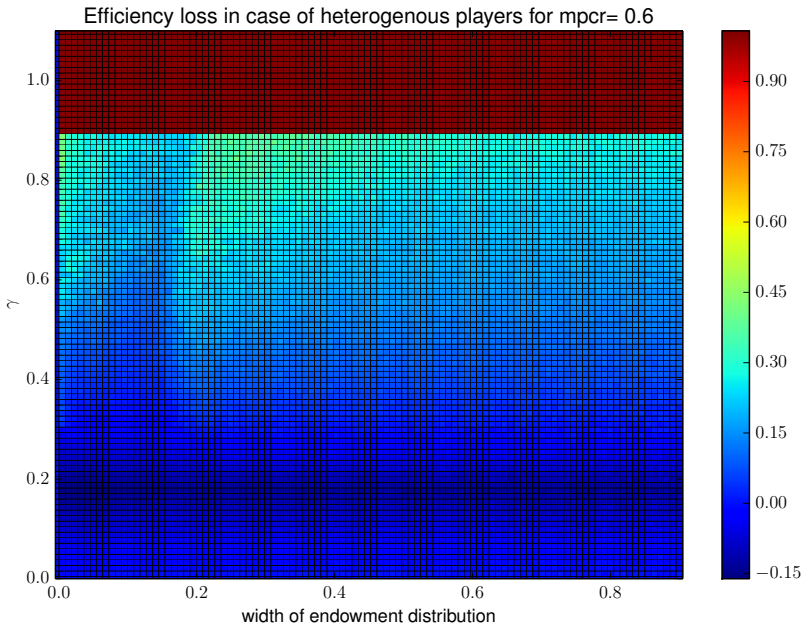


FIGURE A.1: Caption on the following page.

FIGURE A.1: Here we show the loss of efficiency in case of heterogeneous players with respect to the level of contribution that would have been achieved in the homogeneous case. A 100% loss (dark red) indicates that all players contribute nothing and thus that there is a complete loss of efficiency with respect to the homogeneous case. A 0% loss (light blue) means that the system exhibits the same efficiency that it would have with homogeneous players. A negative loss (dark blue) indicates that when endowments are heterogeneous the equilibrium reaches a higher efficiency than it would have in the homogeneous case. As predicted, for superlinear payoffs $\gamma \geq 1$ the only possible equilibrium is non contribution by all and thus the loss of efficiency is total (dark red upper stripe). For intermediate sublinear payoff we can see that the efficiency is not completely lost and that it goes from being quite low to being closer to the homogeneous case. The quantitative value of efficiency that the mixed equilibrium achieves depends on the value of the mpcr and the width of the distribution of initial wealth as well as from other parameters. Finally, for $\gamma \leq \bar{\gamma}$ we first observe a slight increase in efficiency (dark blue line) and then the efficiency approaches the homogeneous one (light blue stripe), on account of the benefits of cooperation being obtainable for an arbitrary small contribution. For the width of the distribution approaching 0 we observe, as expected, the Nash Equilibria in case of homogeneous endowments (blue column on the left). Hence, for a wide range of values of γ , we observe a significant loss in efficiency compared to the homogeneous case. The simulation was obtained for the following set of parameters: $N = 100$, $S = 4$, $Q = 0.6$, $W_0 = 2$, $EN = 50$ and $p = 0.2$. For these parameters, $\bar{\gamma} \approx 0.18$.

The picture shows that the loss of efficiency doesn't seem to change much when changing the width of the endowments distribuion. As expected, for width ≈ 0 , we observe the Nash equilibrium for homogeneous endowments.

For different values of the marginal per capita rate of return, we observe that the higher the mpcr, the wider is the area with partial efficiency losses in the picture and the smaller is the gain in efficiency around $\bar{\gamma}$. The picture below

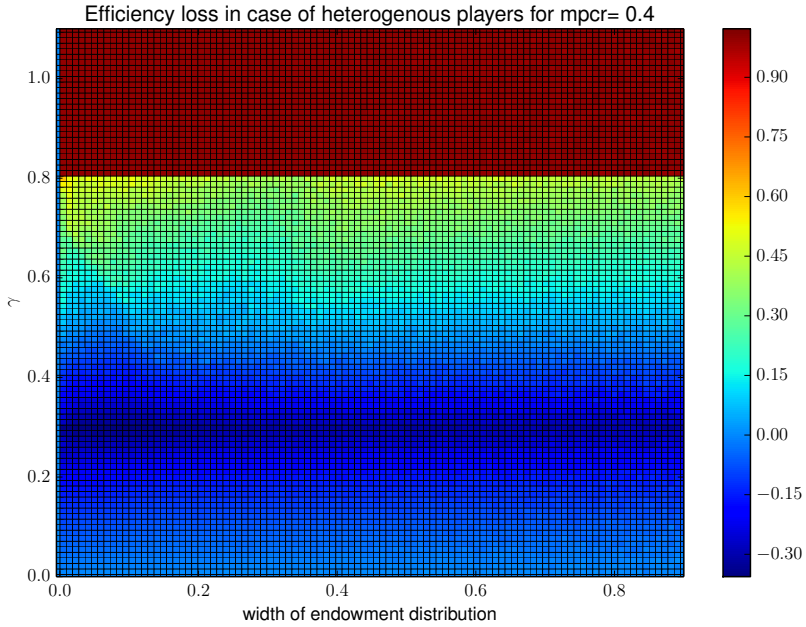


FIGURE A.2: This figure is the equivalent of the previous figure but with a lower marginal per capita rate of return. Compared to figure A.1, we observe a bigger area with complete (100%) efficiency loss. We also observe a much higher increase in the efficiency gain around $\bar{\gamma}$ (which for this parameters is ≈ 0.3).

B

APPENDIX: HETEROGENOUS AGENTS IN VOLUNTARY CONTRIBUTION GAMES WITH ASSORTATIVE MATCHING AND WEALTH ACCUMULATION

B.1 SIMULATION SNAPSHOTS

The following are snapshots of a sample realization of the simulation described in chapter 4 for all the different ranking criteria and taken at an early and later stage of the simulation.

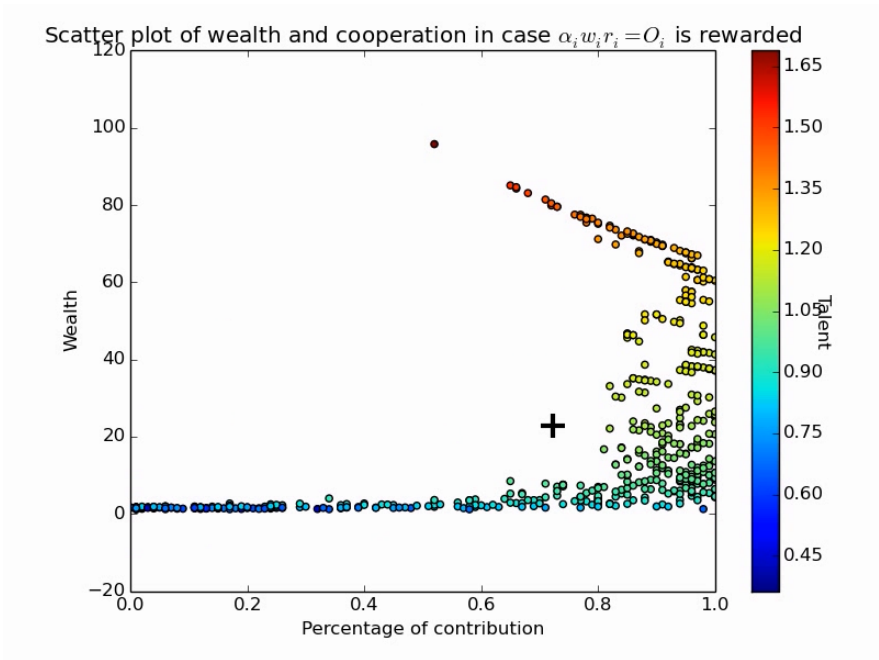


FIGURE B.1: Early snapshot of a sample realization of the simulation described in chapter 4 for the ranking criterion that takes into account the total contributions (case 1). The scatter plot depicts the entire population, with each agent represented by a dot. The abscissa of an agent indicates which percentage of his wealth he contributed to the common pool while his ordinate indicates his current wealth. The black cross marks the population average. The color of each agent indicates how talented he is, ranging from low talent (blue) to high talent (red). The simulation was obtained with the following set of parameters: $N = 500$, $S = 4$, $Q = 0.3$, $\sigma = 0.22$, $\lambda = 20$.

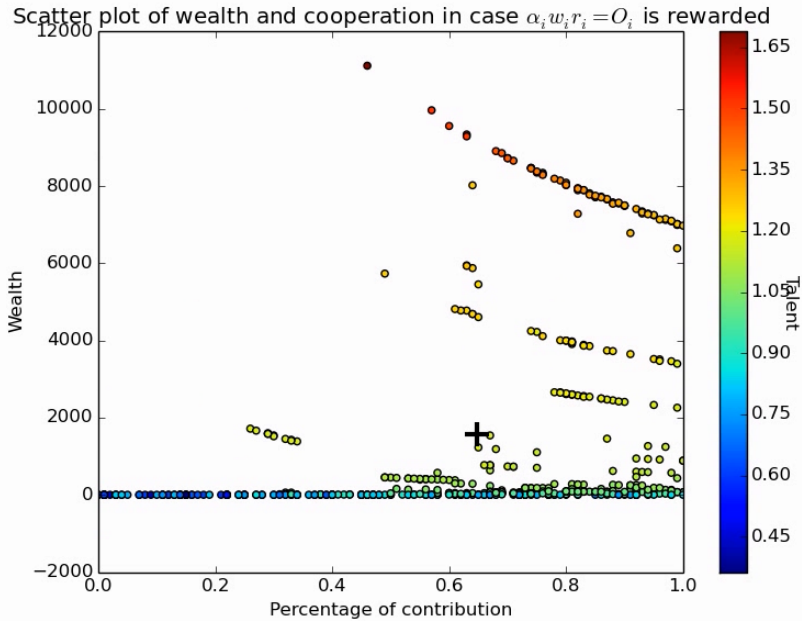


FIGURE B.2: Late snapshot of a sample realization of the simulation described in chapter 4 for the ranking criterion that takes into account the total contributions (case 1). The scatter plot depicts the entire population, with each agent represented by a dot. The abscissa of an agent indicates which percentage of his wealth he contributed to the common pool while his ordinate indicates his current wealth. The black cross marks the population average. The color of each agent indicates how talented he is, ranging from low talent (blue) to high talent (red). The simulation was obtained with the following set of parameters: $N = 500$, $S = 4$, $Q = 0.3$, $\sigma = 0.22$, $\lambda = 20$.

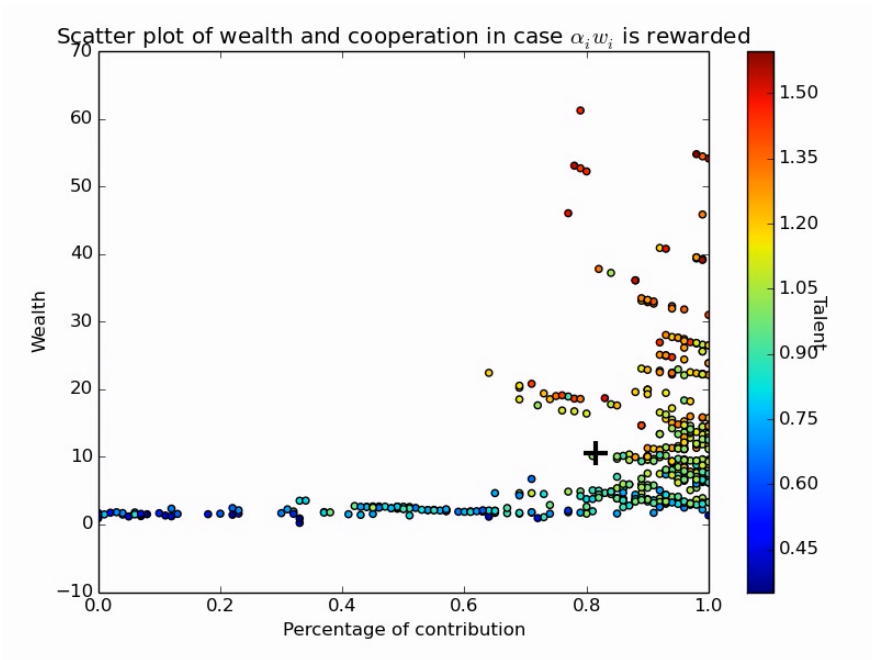


FIGURE B.3: Early snapshot of a sample realization of the simulation described in chapter 4 for the ranking criterion that takes into account only the total wealth contributed by each player (case 2). The scatter plot depicts the entire population, with each agent represented by a dot. The abscissa of an agent indicates which percentage of his wealth he contributed to the common pool while his ordinate indicates his current wealth. The black cross marks the population average. The color of each agent indicates how talented he is, ranging from low talent (blue) to high talent (red). The simulation was obtained with the following set of parameters: $N = 500$, $S = 4$, $Q = 0.3$, $\sigma = 0.22$, $\lambda = 20$.

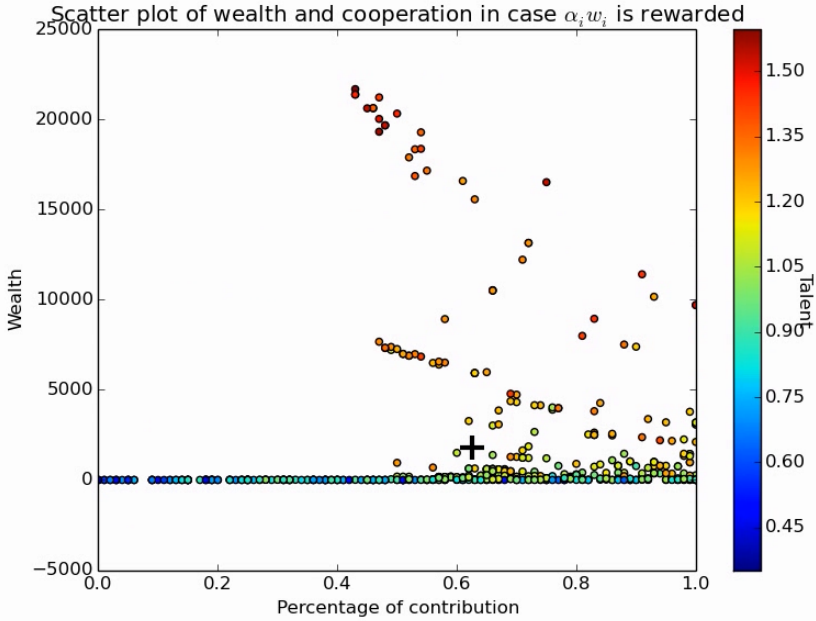


FIGURE B.4: Late snapshot of a sample realization of the simulation described in chapter 4 for the ranking criterion that takes into account only the total wealth contributed by each player (case 2). The scatter plot depicts the entire population, with each agent represented by a dot. The abscissa of an agent indicates which percentage of his wealth he contributed to the common pool while his ordinate indicates his current wealth. The black cross marks the population average. The color of each agent indicates how talented he is, ranging from low talent (blue) to high talent (red). The simulation was obtained with the following set of parameters: $N = 500$, $S = 4$, $Q = 0.3$, $\sigma = 0.22$, $\lambda = 20$.

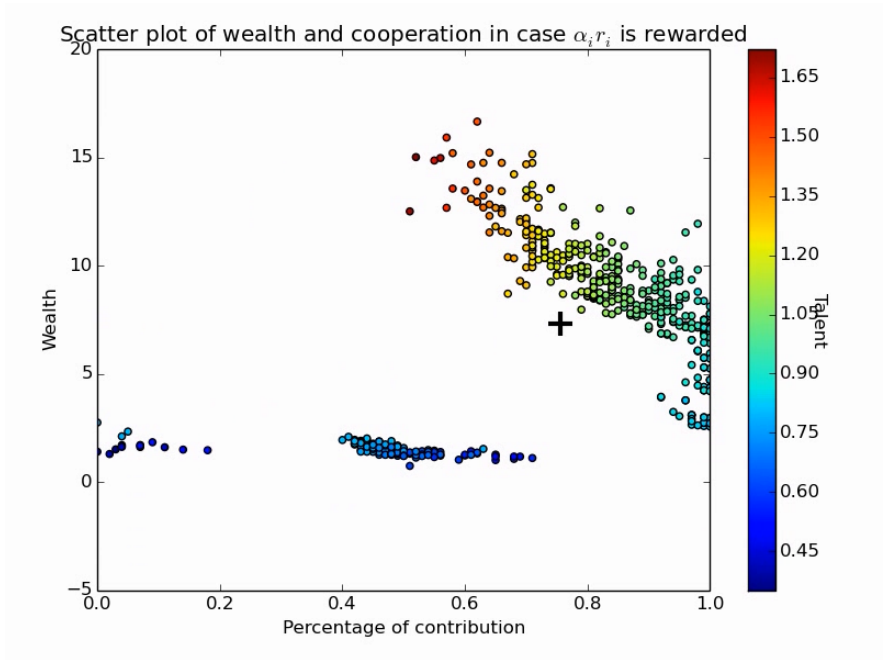


FIGURE B.5: Early snapshot of a sample realization of the simulation described in chapter 4 for the ranking criterion that takes into account the percentage contributed and the talent of each player (case 3). The scatter plot depicts the entire population, with each agent represented by a dot. The abscissa of an agent indicates which percentage of his wealth he contributed to the common pool while his ordinate indicates his current wealth. The black cross marks the population average. The color of each agent indicates how talented he is, ranging from low talent (blue) to high talent (red). The simulation was obtained with the following set of parameters: $N = 500$, $S = 4$, $Q = 0.3$, $\sigma = 0.22$, $\lambda = 20$.

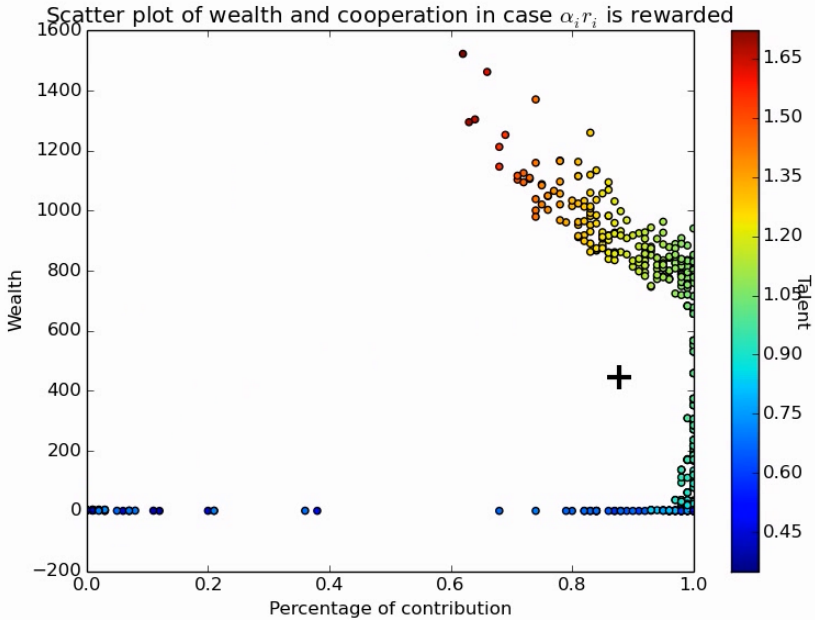


FIGURE B.6: Late snapshot of a sample realization of the simulation described in chapter 4 for the ranking criterion that takes into account the percentage contributed and the talent of each player (case 3). The scatter plot depicts the entire population, with each agent represented by a dot. The abscissa of an agent indicates which percentage of his wealth he contributed to the common pool while his ordinate indicates his current wealth. The black cross marks the population average. The color of each agent indicates how talented he is, ranging from low talent (blue) to high talent (red). The simulation was obtained with the following set of parameters: $N = 500$, $S = 4$, $Q = 0.3$, $\sigma = 0.22$, $\lambda = 20$.

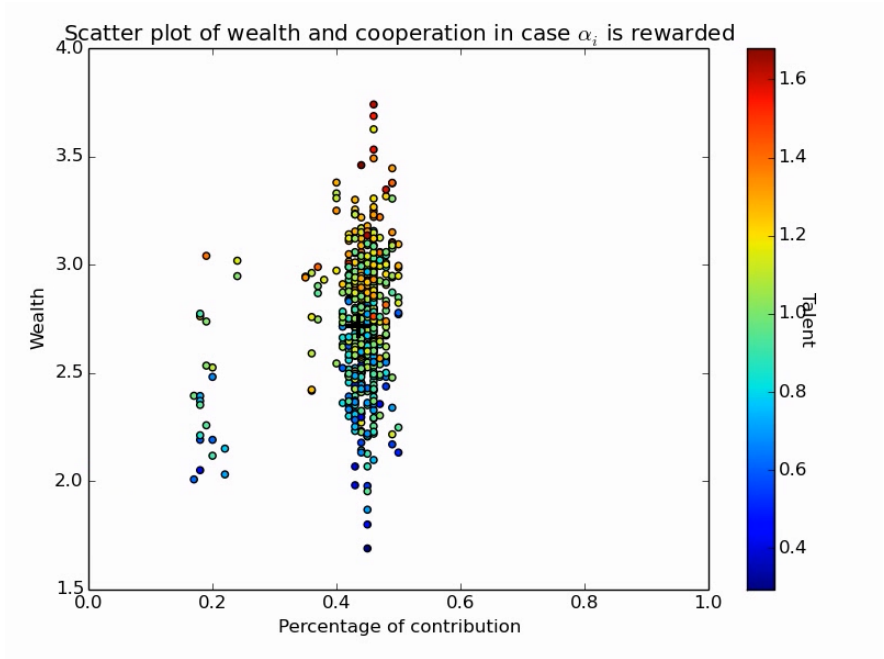


FIGURE B.7: Early snapshot of a sample realization of the simulation described in chapter 4 for the ranking criterion that takes into account only the percentage of wealth contributed by each player (case 4). The scatter plot depicts the entire population, with each agent represented by a dot. The abscissa of an agent indicates which percentage of his wealth he contributed to the common pool while his ordinate indicates his current wealth. The black cross marks the population average. The color of each agent indicates how talented he is, ranging from low talent (blue) to high talent (red). The simulation was obtained with the following set of parameters: $N = 500$, $S = 4$, $Q = 0.3$, $\sigma = 0.22$, $\lambda = 20$.

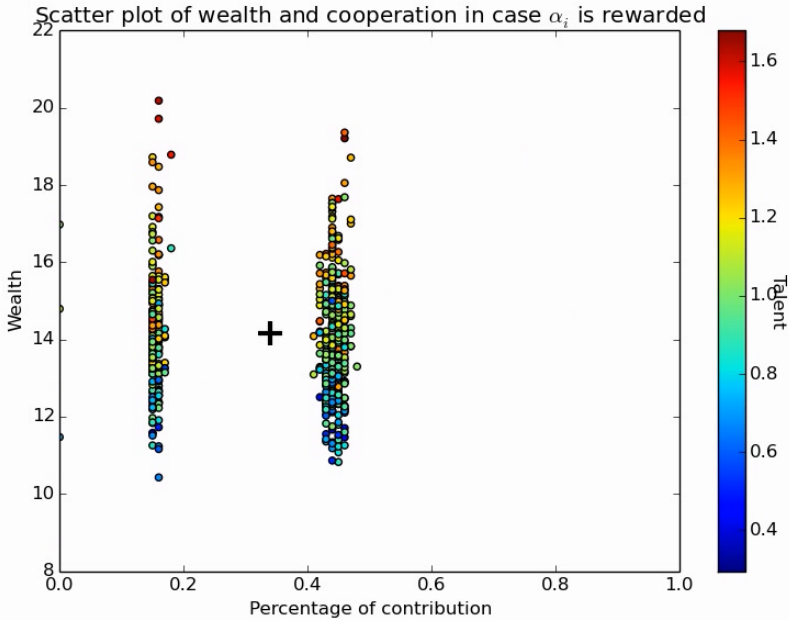


FIGURE B.8: Late snapshot of a sample realization of the simulation described in chapter 4 for the ranking criterion that takes into account only the percentage of wealth contributed by each player (case 4). The scatter plot depicts the entire population, with each agent represented by a dot. The abscissa of an agent indicates which percentage of his wealth he contributed to the common pool while his ordinate indicates his current wealth. The black cross marks the population average. The color of each agent indicates how talented he is, ranging from low talent (blue) to high talent (red). The simulation was obtained with the following set of parameters: $N = 500$, $S = 4$, $Q = 0.3$, $\sigma = 0.22$, $\lambda = 20$.

B.2 SNAPSHOTS OF WEALTH DISTRIBUTION

The following figures show snapshots of the wealth distribution of the population at a late stage of the game for a sample realization of the simulation described in chapter 4 and for all the different ranking criteria. In the pictures, it is clearly possible to observe that in the late stage of the game, most of the wealth produced in the population is owned by very few individuals, with most of the population owning virtually nothing. This seems to hold independently of the ranking criterion used in the simulation, except for when only the percentage of wealth contributed by each player is taken into account (Fig. B.12, case 4).

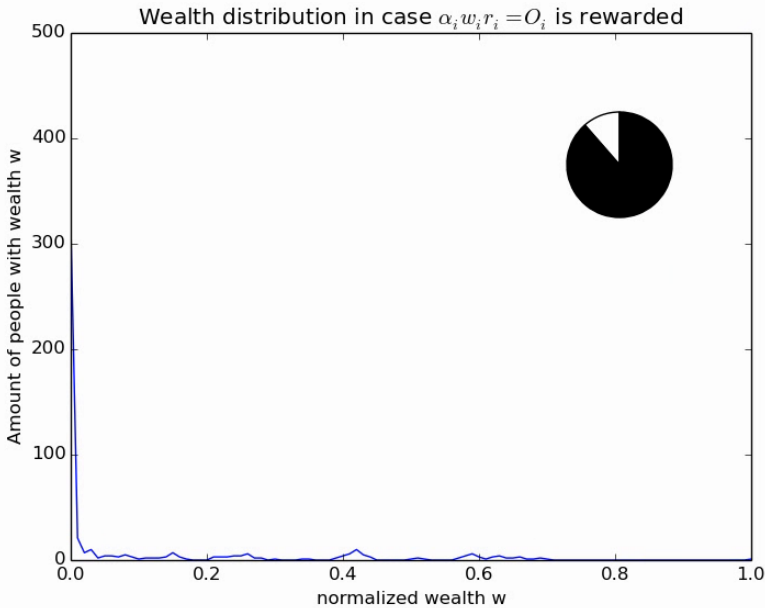


FIGURE B.9: Snapshot of the wealth distribution in the population of a sample realization of the simulation described in chapter 4 for the ranking criterion that takes into account the total contributions (case 1). The line indicates how many agents in the population hold an amount of wealth x , with x being the value of the abscissa. The simulation was obtained with the following set of parameters: $N = 500$, $S = 4$, $Q = 0.3$, $\sigma = 0.22$, $\lambda = 20$.

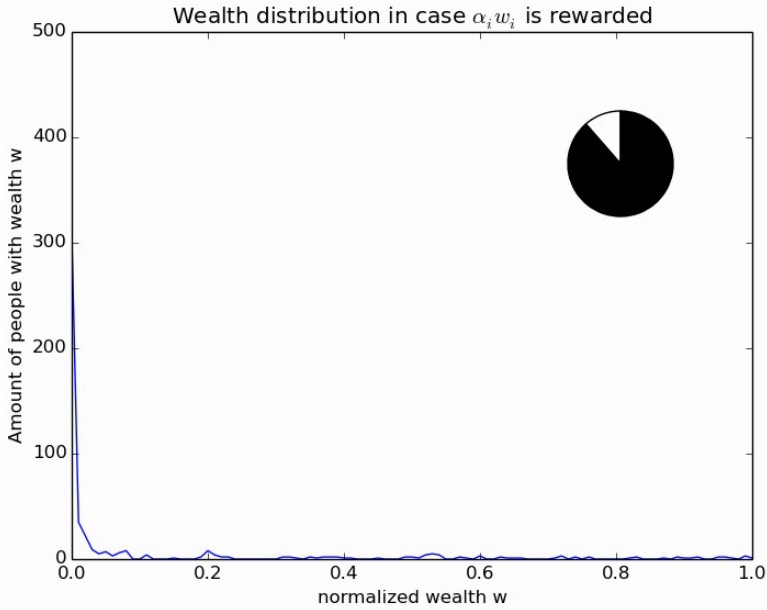


FIGURE B.10: Snapshot of the wealth distribution in the population of a sample realization of the simulation described in chapter 4 for the ranking criterion that takes into account only the total wealth contributed by each player (case 2). The line indicates how many agents in the population hold an amount of wealth x , with x being the value of the abscissa. The simulation was obtained with the following set of parameters: $N = 500$, $S = 4$, $Q = 0.3$, $\sigma = 0.22$, $\lambda = 20$.

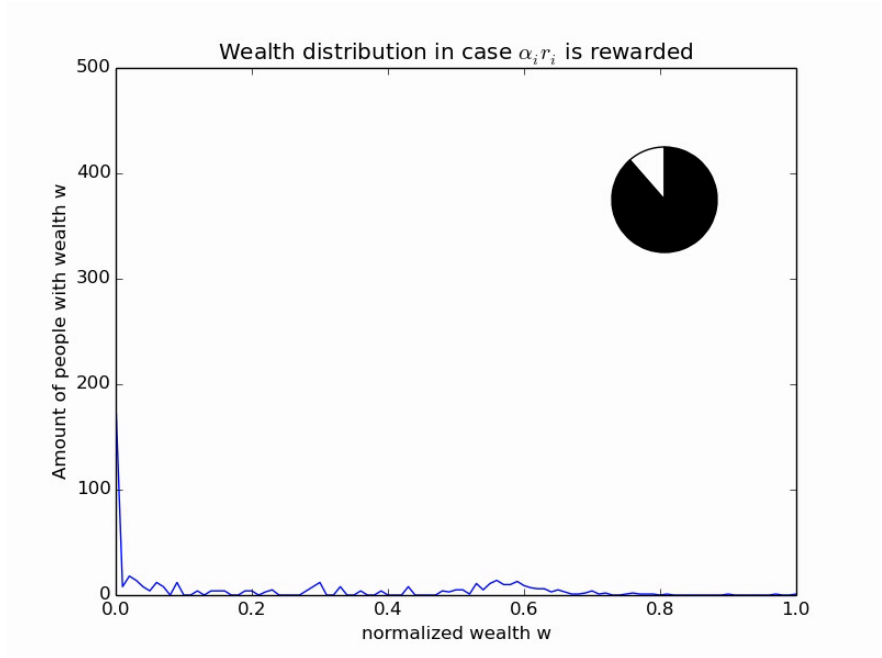


FIGURE B.11: Snapshot of the wealth distribution in the population of a sample realization of the simulation described in chapter 4 for the ranking criterion that takes into account the percentage contributed and the talent of each player (case 3). The line indicates how many agents in the population hold an amount of wealth x , with x being the value of the abscissa. The simulation was obtained with the following set of parameters: $N = 500$, $S = 4$, $Q = 0.3$, $\sigma = 0.22$, $\lambda = 20$.

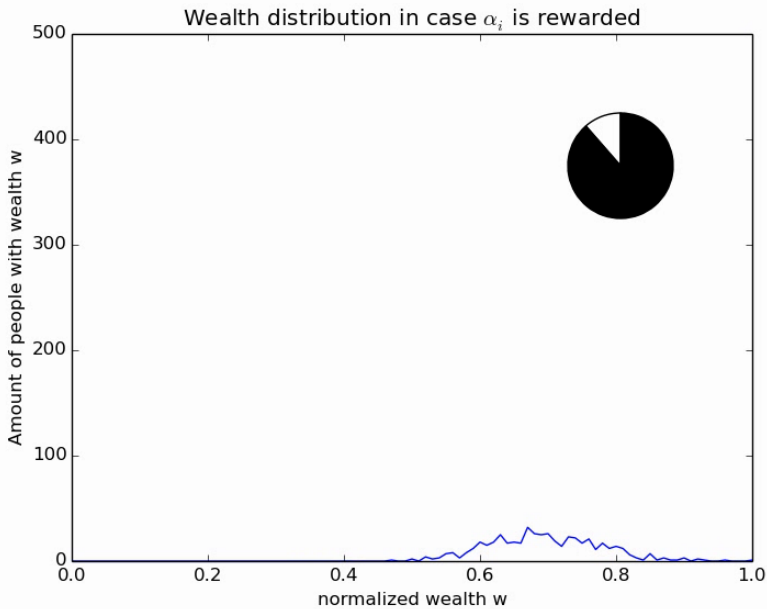


FIGURE B.12: Snapshot of the wealth distribution in the population of a sample realization of the simulation described in chapter 4 for the ranking criterion that takes into account only the percentage of wealth contributed by each player (case 4). The line indicates how many agents in the population hold an amount of wealth x , with x being the value of the abscissa. The simulation was obtained with the following set of parameters: $N = 500$, $S = 4$, $Q = 0.3$, $\sigma = 0.22$, $\lambda = 20$.

B.3 AVERAGE COOPERATION LEVELS

The following figure depicts the average cooperation as a function of time for all the different ranking criteria.

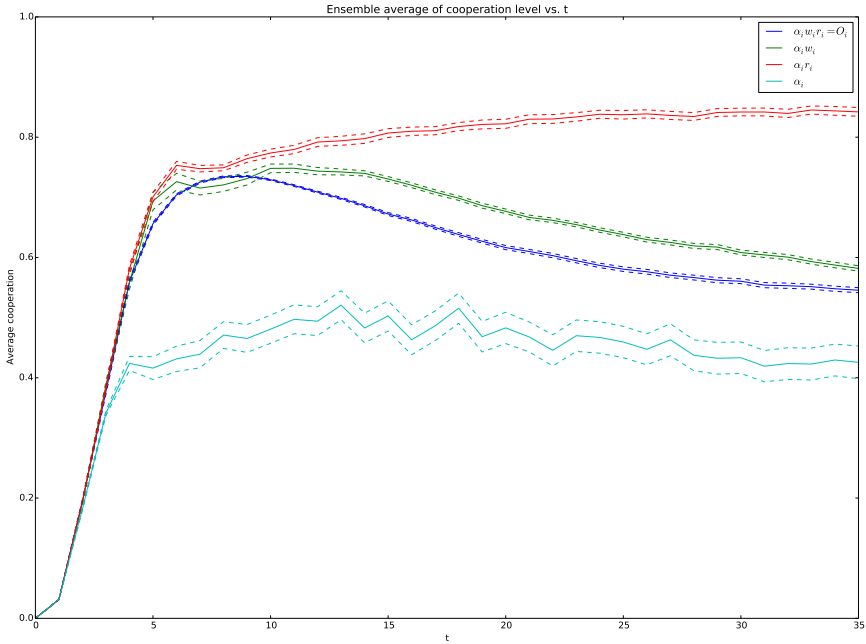


FIGURE B.13: Average cooperation as a function of time for all the ranking criteria: case 1 in blue, case 2 in green, case 3 in red and case 4 in cyan. The results are obtained averaging over 2000 simulations and the dotted lines represent the 95% confidence interval for the ensemble average. After an initial growth in cooperation comparable for all cases, one observes that when agents are ranked solely based on their percentage-wise contribution α_{i_i} , cooperation seems to stabilize around 40% (cyan line). Instead, the ranking criteria that take into account the agents heterogeneity result into a much higher growth for the cooperation (blue, green and red lines). However, when players are ranked taking into account the total amount of wealth contributed, the amount of cooperation in the population starts to decline (blue and green lines). The results were obtained with the following set of parameters: $N = 500$, $S = 4$, $Q = 0.3$, $\sigma = 0.22$, $\lambda = 20$, $T = 350$.

APPENDIX: GROUPS AND SCORES: THE DECLINE OF COOPERATION

C.1 DETAILED EXPERIMENTAL RESULTS

In the following we show the percentage of cooperators as a function of time for each treatment and phase. The plots show data separately for the three phases: the Initial phase (denoted in green) and the first (red) and second (blue) scoring phases. We immediately observe from the pictures that there seem to be no qualitative difference between the behavior of players in the first and second scoring phase.

The errors bars indicate the 95% Binomial proportion confidence interval¹. The difference in the error bars' size depends on the wide difference in the number of available data points².

Because not all treatments were played in every phase (see Table 2 in the main manuscript), not all the plots show data from all the three phases of the experiment; e.g. only the No scoring treatment was played during the initial phase and so for all the other treatments there is no plot for the Initial Phase of the experiment.

¹ For an exact definition see e.g. [186].

² Because we did not run experiments for all possible treatments combinations, the number of data points available for each treatment differs considerably. E.g the Image self scoring treatment was only run before the self scoring treatment, resulting in many less data points for it compared to other treatments. Furthermore, the No scoring treatment has been run at least once in every session (during the first phase), thus resulting in more available data points than all the other treatments.

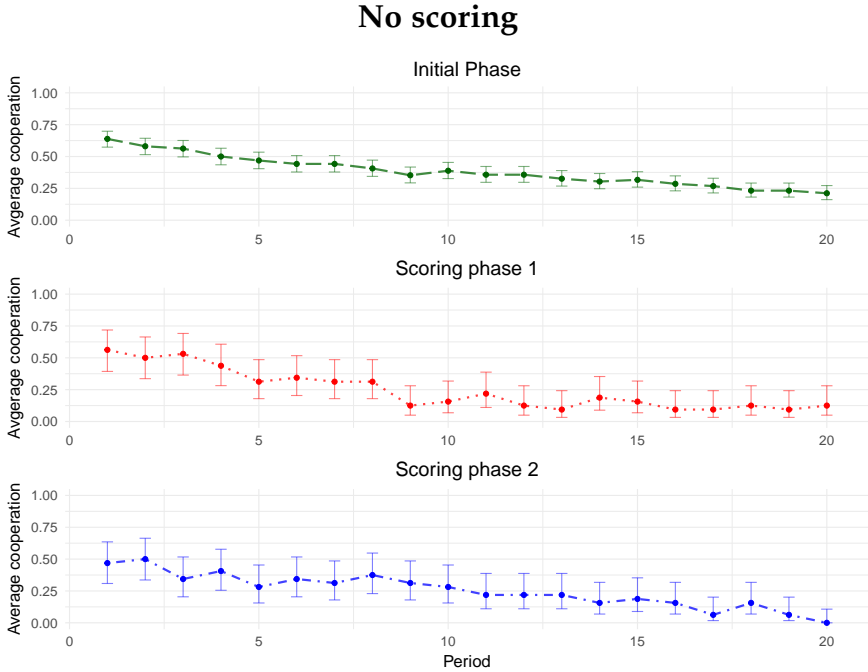


FIGURE C.1: Average percentage of cooperators as a function of time for the No scoring treatment. The three phases are shown separately: the initial phase in green, first scoring phase in red and the second one in blue.

Image scoring

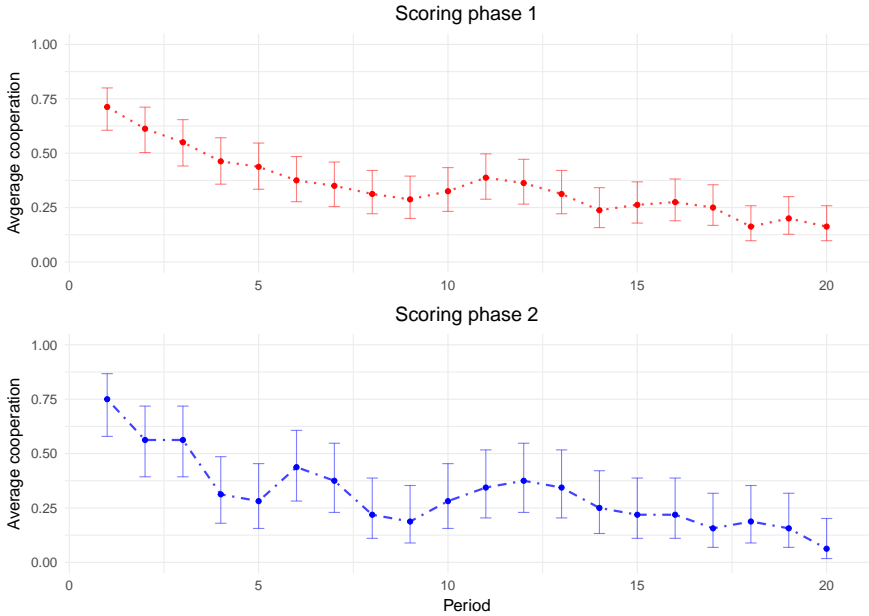


FIGURE C.2: Average percentage of cooperators as a function of time for the Image scoring treatment. The phases are shown separately: the first scoring phase in red and the second one in blue. In no session, the Image scoring treatment was never played during the Initial phase.

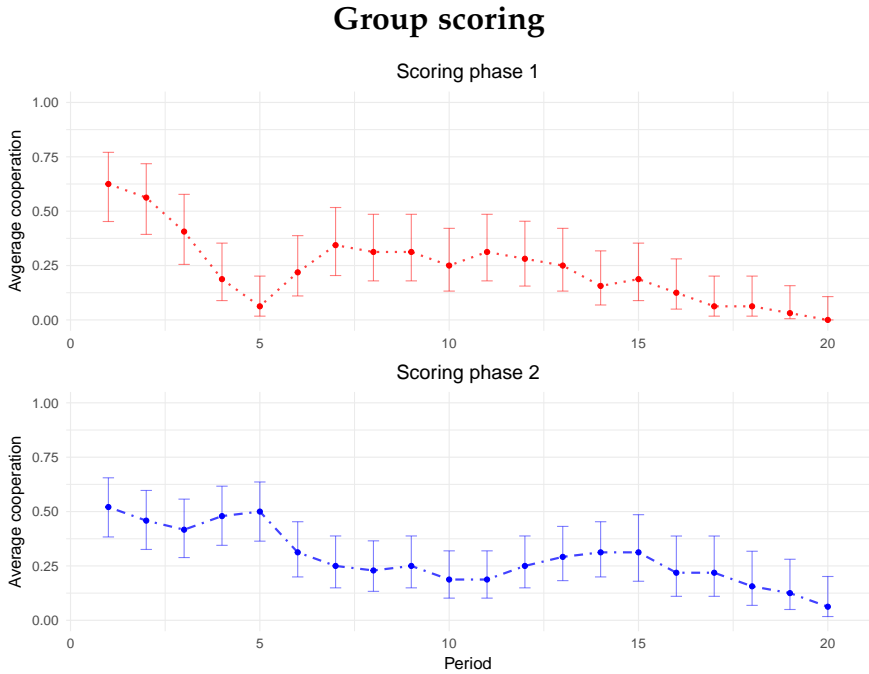


FIGURE C.3: Average percentage of cooperators as a function of time for the Group scoring treatment. The phases are shown separately: the first scoring phase in red and the second one in blue. In no session, the Group scoring treatment was never played during the Initial phase.

Self scoring

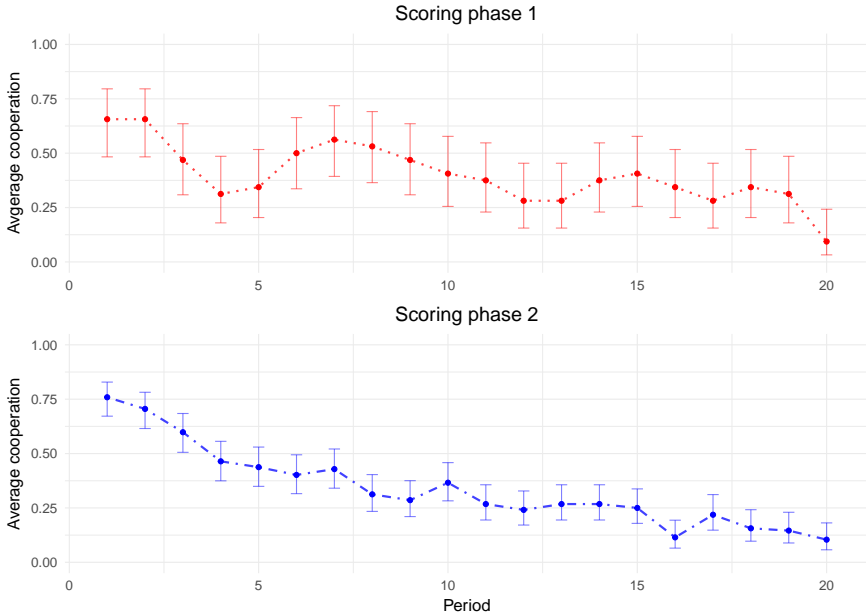


FIGURE C.4: Average percentage of cooperators as a function of time for the Self scoring treatment. The phases are shown separately: the first scoring phase in red and the second one in blue. In no session, the Image scoring treatment was never played during the Self phase.

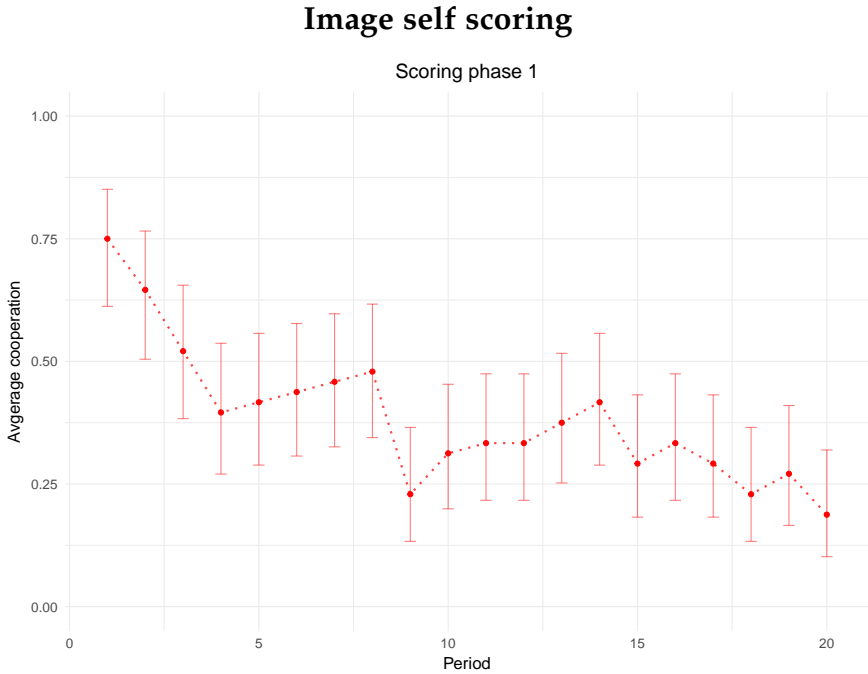


FIGURE C.5: Average percentage of cooperators as a function of time for the Image self scoring treatment. The Image self scoring treatment was only played during the first scoring phase.

C.2 PERCENTAGE OF COOPERATORS AS A FUNCTION OF THE OBSERVED SCORE IN THEIR GROUP

In the following, we plot the percentage of cooperators as a function of the observed score in their group.

In the Image and Group score figures, we can observe that players contribute more with increasing observed score in their group. However, players seem to do so with a downward bias, especially for high score values; this downward bias results in a steady reduction of "good players"³ in the population and thus in the breakdown of whatever positive effect the scoring mechanism might have had on the cooperative behavior of players.

In the plots below, the bars indicate the 95% Binomial proportion confidence interval, computed using the Wilson score interval⁴. The Wilson score was chosen because it is well behaved even for small sample sizes and extreme probabilities, and it returns asymmetric confidence intervals⁴.

The numbers at the side of each point indicate how many experimental points contributed to the average. The different number of observations derives from the fact that different scoring mechanisms produce a different number of possible combinations of scores in a group.

³ I.e. players with a high score.

⁴ A focal property of the Wilson score is that it returns non-vanishing error bars for small sample sizes even when the expected probability is 0 or 1.

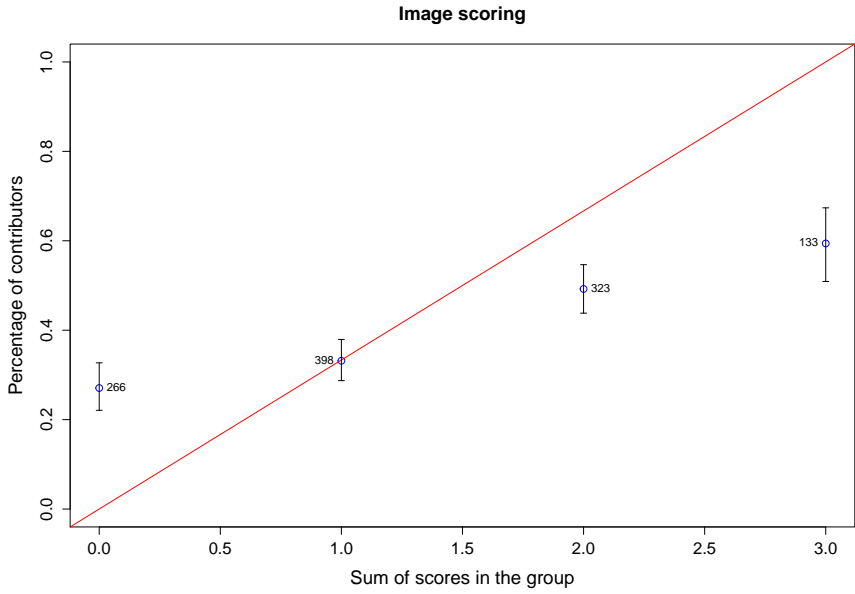


FIGURE C.6: Cooperation as a function of observed score for the Image scoring treatment. The red line in the figure represents the identity line.

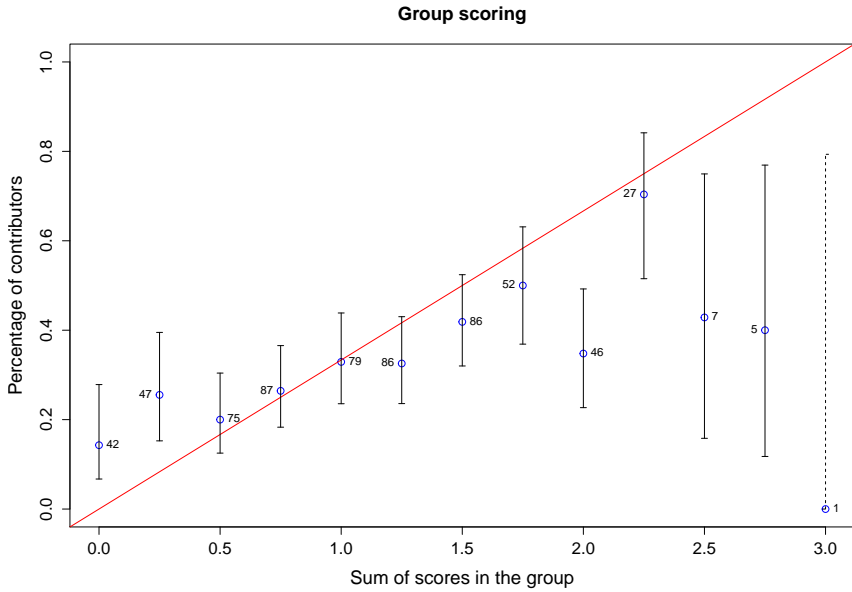


FIGURE C.7: Cooperation as a function of observed score for the Group scoring treatment. The red line in the figure represents the identity line. The dashed error bar indicates that, due to the low number of observations for that point, it is not possible to provide a good estimation of the confidence interval; nevertheless, the best estimation is provided.

Due to the high numbers of possible values for the Self and Image Self scores, the related figures are more chaotic but a similar trend seems to be present as well.

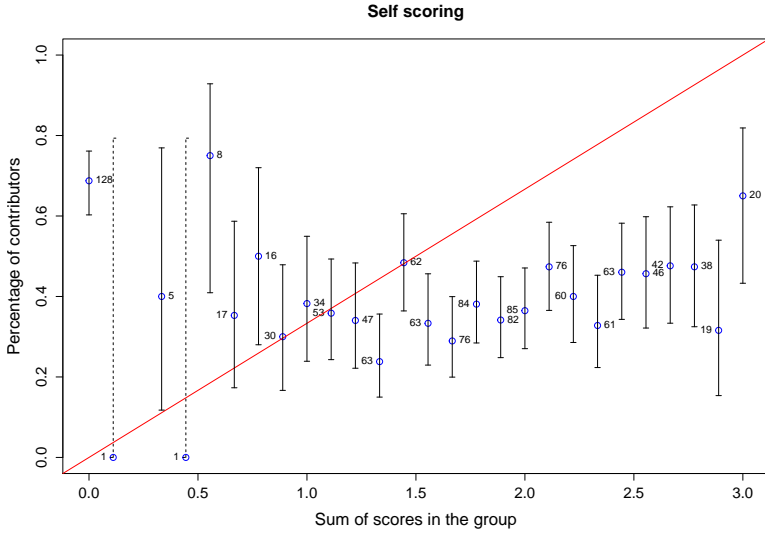


FIGURE C.8: Cooperation as a function of observed score for the Self scoring treatment. The red line in the figure represents the identity line. The dashed error bars indicate that, due to the low number of observations for that points, it is not possible to provide a good estimation of the confidence intervals; nevertheless, the best estimations are provided.

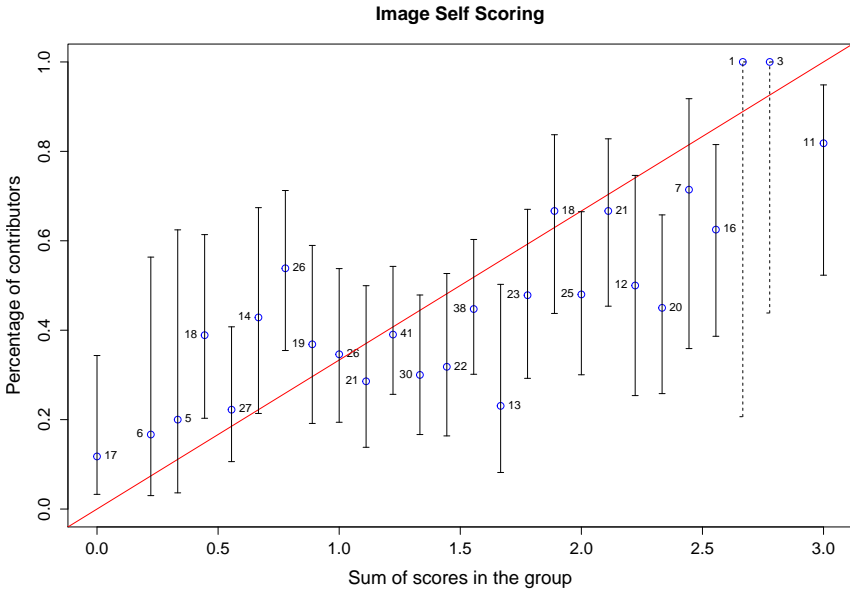


FIGURE C.9: Cooperation as a function of observed score for the Image self scoring treatment. The red line in the figure represents the identity line. The dashed error bars indicate that, due to the low number of observations for that points, it is not possible to provide a good estimation of the confidence intervals; nevertheless, the best estimations are provided.

We now look at the case where no score at all was provided to the players: in the following we plot the percentage of cooperation in the group as a function of the image score that the players *would* have observed *if* the information would have been provided, i.e. as a function of people in the group that contributed in the previous round. In this case we observe no apparent trend of increasing cooperation.

This suggests that players were indeed responding to the observed scores in the group, and that higher levels of cooperation were not just an artifact due to the correlation between the high scores and high levels of cooperation, i.e. purely a result of conditional cooperation.

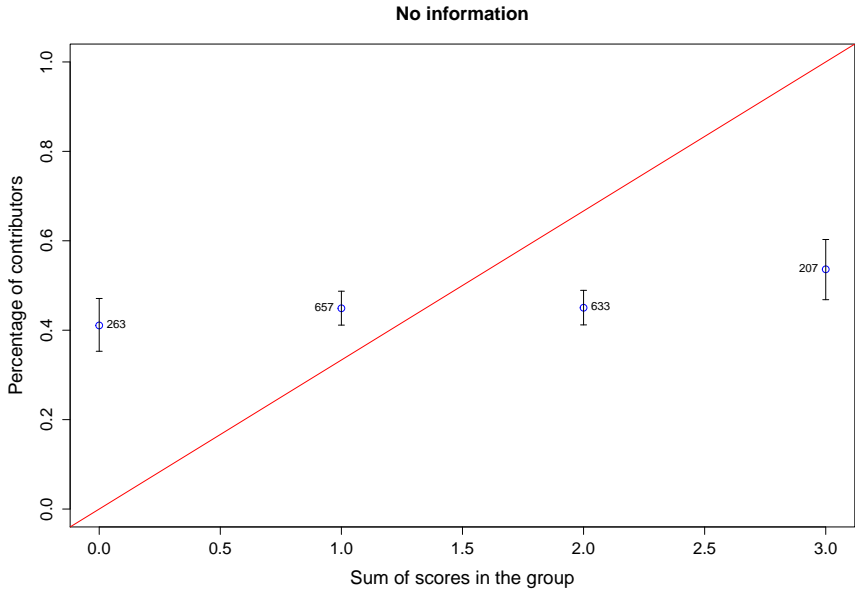


FIGURE C.10: Cooperation as a function of what *would have been* the observed score for the No scoring treatment. The red line in the figure represents the identity line.

In order to further highlight the difference between the cases where scoring is provided or not, we also plot together the image scoring (in black) and no scoring case (in black).

Comparison between Image and No scoring

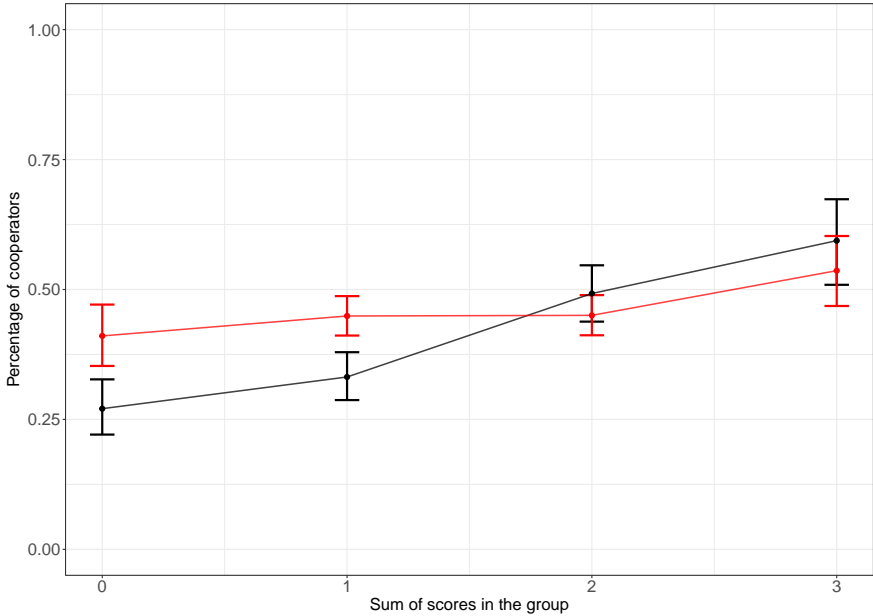


FIGURE C.11: Comparison between the No scoring (black) and the Image scoring (red) treatments: The above picture shows the average cooperation as a function of the scores in the group. We can discern an upward trend for the Image scoring case (red line) but there seem to be no significant trend in the case of No scoring (black line). This suggests that players are indeed responding to other players' scores.

C.3 FURTHER STATISTICAL ANALYSIS

In the following, we present a different statistical analysis to test whether it is possible to statistically distinguish between treatments.

In accordance to the test presented in the main manuscript, because of the potential correlations between decisions taken by player i at time t and all decisions taken at $t - 1$ (also known as temporal autocorrelations), we decided to perform the statistical test only on the data acquired during the first round of each phase. For completeness, at the end of this section we present the same test performed on the entire dataset.

Our goal is to try to understand whether the null hypothesis H_0 that treatments have no significant impact on the decisions taken by the agents is verified. For this reason, we performed a permutation test⁵ using the between sum of squares (SSB) as statistics:

$$SSB = \sum_{i=1}^m n_i (\bar{y} - \bar{y}_i)^2$$

where i represents a treatment, \bar{y}_i the average cooperation in that treatment and \bar{y} the overall average cooperation.

The idea behind the test is to randomly permute treatments within sessions many times, thus creating a random sample of all the possible matches between our observable and the treatment under which they have been observed (see figure for a two dimensional example with 3 treatments tested 3 times each).

5 See e.g. [187–190] for a review and example applications of permutation tests.

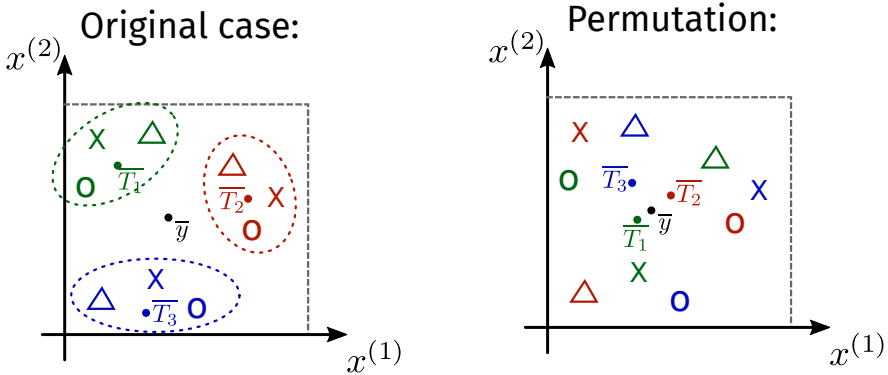


FIGURE C.12: Example of permutation. Every symbol represents data collected in an experiment session. X , Δ and O indicate that the data were collected during the first, second and third session respectively. **Green** indicates that data were collected during treatment 1, **red** during treatment 2 and **blue** during treatment 3. \bar{T}_i is the mean for treatment i while \bar{y} represents the global mean.

For every random permutation we computed the resulting SSB, thus obtaining an histogram representing the empirical distribution of the SSB values.

We can now compare two hypothesis: The null hypothesis (H_0), where treatments have no impact on the decisions taken by the agents and the alternative hypothesis (H_1) where we should observe statistical significance of treatments.

To compare the two hypothesis, we calculate the SBB for our original data (let us call it SBB^*) and we compute how likely it is (using the empirical observed cumulative distribution function) to observe SBB^* under H_0 .

Here is what we obtain:

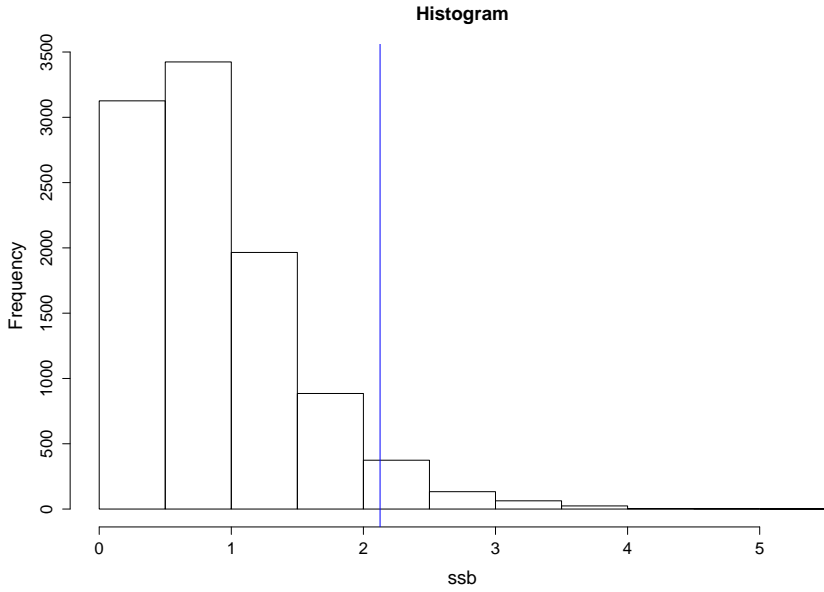


FIGURE C.13: The histogram represents the frequency of observed sbb values obtained from 1000000 permutations. The blue line indicates where the SBB for the original data lies.

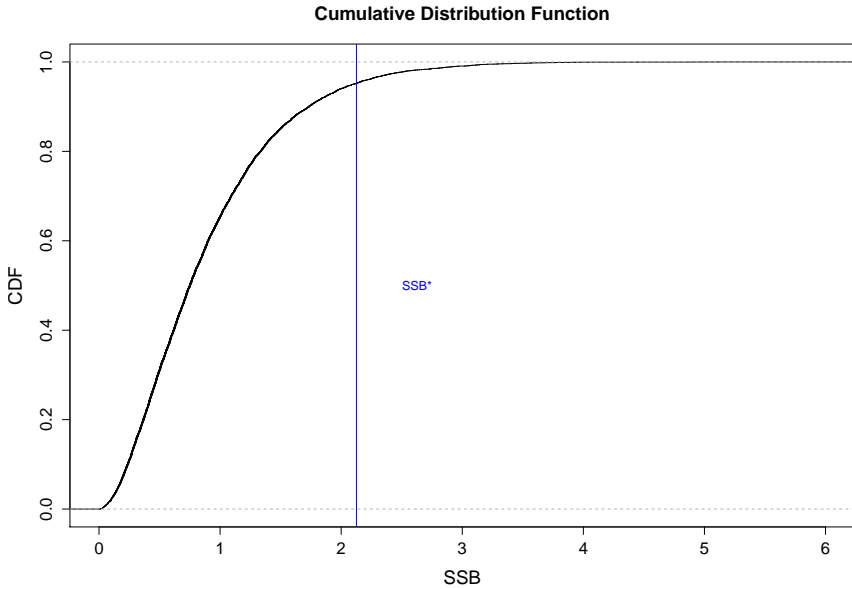


FIGURE C.14: Here we plot the empirical cumulative distribution function obtained from the permutation. From it, it is possible to compute the empirical p-value for SSB^* .

From the figures above we can clearly observe that H_0 is rejected and thus that the treatments are not all statistically indistinguishable from each other (the observed p-value is < 0.05). This is in line with the results presented in table 3 where we observe that some treatments are significantly different from each other.

For completeness, in the following we show the histogram and empirical CDF for the entire dataset (note that some of the points in the complete dataset may be autocorrelated):

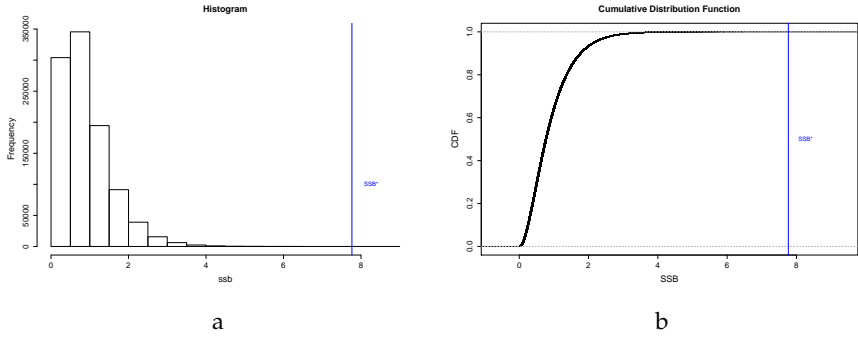


FIGURE C.15: Histogram (a) and empirical CDF (b) obtained using the complete dataset.

C.4 DECISION MAKING MICRO MODEL

In the following we provide a simple model for each player's decision making and we subsequently fit it to our data.

We hypothesize that a player could act according to one or more of these behavioral rules:

- **Unconditional defection (cooperation):** Players could just unconditionally defect or cooperate, regardless of the other players' actions. We find that slightly less of 20% of the subjects in our dataset are unconditional defectors and that two players are unconditional cooperators.
- **Conditional cooperation:** Players' experiences in their previous group(s) might influence their propensity to cooperate in the future. This implies that players are sensitive to the decisions taken by other players in the past so that the more players cooperated with them in the past, the more will they tend to cooperate in the future⁶. Due to the binary nature of actions in the multi-player PD, we get non linear thresholds policies for conditional cooperation; i.e. a player might decide to cooperate if 1,2 or 3 players (other than himself) cooperated in his previous group. Figure C.16 shows how a conditional cooperator would behave depending on the number of people N that cooperated in his previous group and the conditional cooperation threshold.
- **Indirect reciprocity:** Players have a propensity to cooperate with people that they know cooperated with others in the previous interaction. This means that players are sensitive to the score of the players with which they interact. Again, the binary nature of actions results in more than threshold also for indirect reciprocity; i.e. players might decide to cooperate if the aggregated score that they observe in the group is equal or higher than different values (fixed in our case to 1,2 or 3). Figure C.16 shows how an indirect reciprocator would behave depending on the observed sum of scores and his indirect reciprocity threshold.
- **Learning:** While playing, players might figure out which actions lead to higher payoff. Hence a player would decide to keep his current strategy if the payoff received in the current round is at least equal to the one received in the previous round and switch action otherwise.

⁶ See e.g. [44, 47, 191] for theoretical and empirical results on conditional cooperation.

Figure C.17 shows how a player behaving according to our definition of learning would choose his following action based on his actions in the current and previous rounds (respectively a_t and a_{t-1}) and the resulting payoffs (respectively ϕ_t and ϕ_{t-1}).

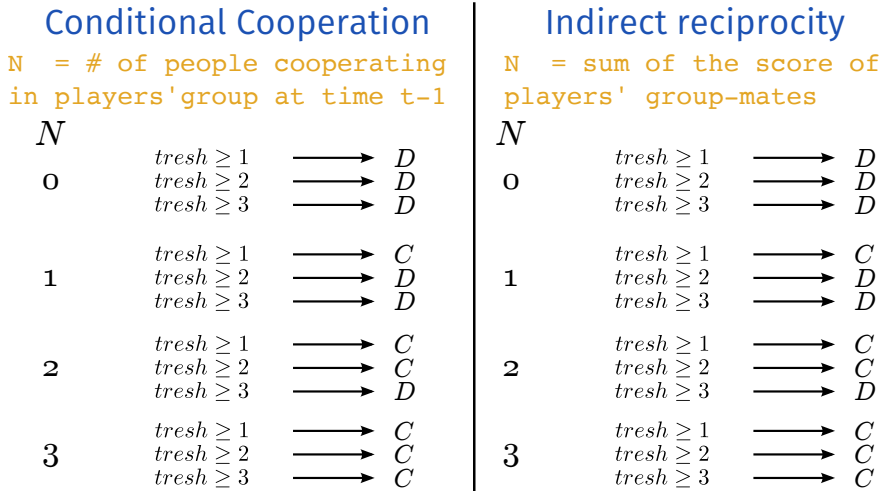


FIGURE C.16: Predicted action for conditional cooperation (left) and indirect reciprocity (right). Due to the binary nature of actions in the multi-player PD, we obtain non linear thresholds policies for both conditional cooperation and indirect reciprocity. The decision to cooperate or not for a conditional cooperator will depend on the number of people N that cooperated in his previous group and on his conditional cooperation threshold. The decision to cooperate or not for an indirect reciprocator depends on the sum of the scores observed in his current group N and on his indirect reciprocity threshold (here set at 1,2 or 3).

We fitted these behavioral rules to the data collected in our experiment and the results are in line with the macroscopic analysis presented in the main manuscript: We find that people significantly react to the information provided by the score. In fact, people seem to behave more according to indirect reciprocity rules, the more robust (or trustworthy) is the scoring mechanism; e.g. in the image scoring treatments players seem to rely more on an indirect reciprocity strategy than in the group scoring treatments and even more so for the self scoring treatment. When information on the

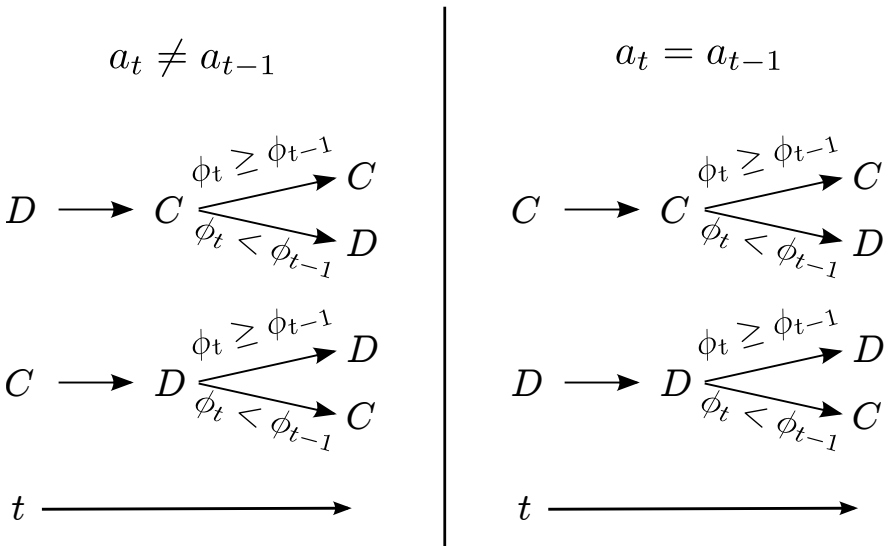


FIGURE C.17: Predicted action for a player behaving according to (our definition) of learning. A player would decide to keep his current action ($a_t \in \{C, D\}$) if the payoff received in the current round ϕ_t is at least equal to the one received in the previous round ϕ_{t-1} and switch action otherwise.

score of players is either unreliable (as in the self scoring treatment) or unavailable (as in the no scoring case), people revert a conditional cooperation behavior.

Crucially, we find that players seem to have high thresholds for both conditional cooperation and indirect reciprocity. This implies that players react to other players cooperating, but not as much as they should to keep a steady percentage of cooperative players in the population. This is in perfect accord with what observed in Figure 3 in the main manuscript and it results in a steady shrinking of players with a positive score, thus negating the positive effect of reputation on cooperative behavior culminating in the "spiraling down" of cooperation.

We also find that there is a significant percentage of learning behavior across all treatments. Table C.1 summarizes our findings.

For a more detailed explanation of the regressions that we performed and for detailed results, we refer the readers to the following subsections: in subsection C.4.1 we present the detailed results of regressions using the first 12 out of 20 periods and in subsection C.4.2 the results using all periods. Finally, in subsection C.4.3 we show the marginal effect of each independent variable over the dependent variable.

	Learning	Cond. Coop.	Indirect Rec.
No information	✓✓	✓	NA
Image scoring	✓✓	✗	✓✓
Group scoring	✓	✗	✓
Self scoring	✓✓	✓✓	✗
Image self scoring	✓	✗	✓

TABLE C.1: Summary of the outcome of the regressions: In the above table we summarize how significant for each treatment are Conditional Cooperation, Indirect Reciprocity and Learning. A double tick indicates that a behavioral rule is highly significant in a given treatment, a single tick signals that a rule is somehow significant and a cross indicates no evidence of that behavior in a given treatment. The results shown in the table are in agreement with the macro analysis presented in the main text: people seem to behave more according to indirect reciprocity rules, the more robust is the scoring mechanism. When scores are unreliable or not available, players revert to conditional cooperation. We also find for all treatments a significant percentage of learning behavior.

C.4.1 Fit for first 12 rounds

Due to the large decline in cooperation happening during the late rounds of every treatment, it is harder to distinguish between different behavioral rules since all of the rules we consider would predict the same outcome when nearly everyone defects. For this reason, we performed our regressions using only the first 12 rounds of each phase, thus resulting in 10 data points for each phase⁷. Regressions taking into account the full dataset, are presented in the next subsection.

For every treatment, we performed a linear and a probit regression of the full model and of subsets of it.

The full model with which we fit our data is:

$$\begin{aligned} \text{Action} = & \alpha + \beta_1 \cdot \text{OneCC} + \beta_2 \cdot \text{TwoCC} + \beta_3 \cdot \text{ThreeCC} + \beta_4 \cdot \text{OneIR} \\ & + \beta_5 \cdot \text{TwoIR} + \beta_6 \cdot \text{ThreeIR} + \beta_7 \cdot \text{Learn_Inert} + \beta_8 \cdot \text{tindex} \quad (\text{C.1}) \end{aligned}$$

where

⁷ The predictions of the behavioral rules can be tested starting from round 3, because one needs a_t and a_{t-1} to predict the action at a_{t+1} .

- **Action** is the dependent variable. It can only take two values: either defect or cooperate.
- **OneCC** is a dummy variable signaling whether at least one person cooperated in the group of the active player in the previous round.
- **TwoCC** is a dummy variable signaling whether at least two people cooperated in the group of the active player in the previous round.
- **ThreeCC** is a dummy variable signaling whether at least three people cooperated in the group of the active player in the previous round.
- **OneIR** is a dummy variable signaling whether the sum of the scores in the active player's group (excluding the active player's score) is at least one.
- **TwoIR** is a dummy variable signaling whether the sum of the scores in the active player's group (excluding the active player's score) is at least two.
- **ThreeIR** is a dummy variable signaling whether the sum of the scores in the active player's group (excluding the active player's score) is at least three.
- **LearnInert** is a dummy variable taking signaling whether or not the player's choice is consistent with the learning behavior (as defined above).
- **tindex** represents the round in which the decision takes place. It ranges from 3 to 12.

Below are the results of the regressions that we performed, presented using the *texreg* [192] package. For every treatment we performed a linear and a probit regression for the entire model and then a probit regression for subsets of the model, i.e. taking into account only conditional cooperation or indirect reciprocity (OnlyCC & OnlyIR) or only one of the three possible thresholds (Only1, Only2, Only3).

No scoring

	Probit	Linear	OnlyCC	Only1	Only2	Only3
(Intercept)	-0.12 (0.09)	0.45*** (0.03)	-0.12 (0.09)	-0.09 (0.08)	0.04 (0.07)	0.10 (0.06)
OneCC	0.18** (0.06)	0.06** (0.02)	0.18** (0.06)	0.22*** (0.06)		
TwoCC	0.08 (0.05)	0.03 (0.02)	0.08 (0.05)		0.16*** (0.05)	
ThreeCC	0.22* (0.09)	0.08* (0.03)	0.22* (0.09)			0.30*** (0.09)
Learn_Inert	0.37*** (0.04)	0.14*** (0.02)	0.37*** (0.04)	0.38*** (0.04)	0.37*** (0.04)	0.37*** (0.04)
tindex	-0.06*** (0.01)	-0.02*** (0.00)	-0.06*** (0.01)	-0.07*** (0.01)	-0.07*** (0.01)	-0.07*** (0.01)
AIC	4786.39		4786.39	4793.83	4797.24	4798.37
BIC	4823.73		4823.73	4818.72	4822.13	4823.26
Log Likelihood	-2387.19		-2387.19	-2392.91	-2394.62	-2395.18
Deviance	4774.39		4774.39	4785.83	4789.24	4790.37
Num. obs.	3726	3726	3726	3726	3726	3726
R ²		0.06				
Adj. R ²		0.06				
RMSE		0.47				

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

TABLE C.2: Regression results for the No scoring treatment when only the first 12 rounds are taken into account. Due to the lack of information about the scores of players in this treatment, the model considered here doesn't contain the indirect reciprocity dummy variables.

Image scoring

	Probit	Linear	OnlyCC	OnlyIR	Only1	Only2	Only3
(Intercept)	-0.76*** (0.18)	0.22*** (0.06)	-0.12 (0.15)	-0.94*** (0.16)	-0.56** (0.17)	-0.78*** (0.15)	-0.50*** (0.13)
OneCC	-0.23* (0.10)	-0.07* (0.04)	-0.12 (0.10)		-0.10 (0.09)		
TwoCC	-0.02 (0.11)	-0.00 (0.04)	0.12 (0.10)			-0.04 (0.09)	
ThreeCC	0.08 (0.16)	0.02 (0.05)	0.27 (0.15)				0.11 (0.14)
OneIR	0.14 (0.13)	0.04 (0.04)		0.12 (0.12)	0.60*** (0.11)		
TwoIR	0.65*** (0.11)	0.23*** (0.04)		0.63*** (0.10)		0.82*** (0.09)	
ThreeIR	0.42*** (0.12)	0.16*** (0.04)		0.40*** (0.12)			0.76*** (0.11)
Learn_Inert	0.42*** (0.08)	0.14*** (0.03)	0.45*** (0.08)	0.44*** (0.08)	0.45*** (0.08)	0.45*** (0.08)	0.44*** (0.08)
tindex	-0.01 (0.02)	-0.00 (0.01)	-0.04** (0.01)	-0.00 (0.02)	-0.04** (0.01)	-0.01 (0.01)	-0.02 (0.01)
AIC	1275.68		1376.72	1275.54	1350.92	1285.87	1326.07
BIC	1320.38		1406.52	1305.34	1375.75	1310.70	1350.90
Log Likelihood	-628.84		-682.36	-631.77	-670.46	-637.93	-658.03
Deviance	1257.68		1364.72	1263.54	1340.92	1275.87	1316.07
Num. obs.	1060	1060	1060	1060	1060	1060	1060
R ²		0.15					
Adj. R ²		0.14					
RMSE		0.45					

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

TABLE C.3: Regression results for the Image scoring treatment when only the first 12 rounds are taken into account.

Group scoring

	Probit	Linear	OnlyCC	OnlyIR	Only1	Only2	Only3
(Intercept)	-0.81*** (0.19)	0.21*** (0.06)	-0.54** (0.17)	-0.79*** (0.18)	-0.77*** (0.19)	-0.55*** (0.16)	-0.30* (0.14)
OneCC	0.09 (0.12)	0.03 (0.04)	0.23* (0.11)		0.13 (0.11)		
TwoCC	-0.07 (0.14)	-0.02 (0.05)	0.12 (0.13)			-0.02 (0.12)	
ThreeCC	-0.23 (0.25)	-0.09 (0.09)	-0.02 (0.24)				-0.01 (0.23)
OneIR	0.26* (0.12)	0.08* (0.04)		0.28* (0.11)	0.38*** (0.11)		
TwoIR	0.49*** (0.14)	0.19*** (0.05)		0.46*** (0.13)		0.61*** (0.13)	
ThreeIR	0.51 (0.40)	0.20 (0.14)		0.42 (0.39)			0.82* (0.39)
Learn_Inert	0.26* (0.10)	0.09* (0.03)	0.30** (0.10)	0.25* (0.10)	0.29** (0.10)	0.25* (0.10)	0.27** (0.10)
tindex	-0.02 (0.02)	-0.01 (0.01)	-0.04* (0.02)	-0.01 (0.02)	-0.02 (0.02)	-0.02 (0.02)	-0.04* (0.02)
AIC	910.91		931.29	906.62	919.01	911.74	932.04
BIC	952.73		959.17	934.50	942.24	934.97	955.27
Log Likelihood	-446.46		-459.65	-447.31	-454.50	-450.87	-461.02
Deviance	892.91		919.29	894.62	909.01	901.74	922.04
Num. obs.	770	770	770	770	770	770	770
R ²		0.07					
Adj. R ²		0.06					
RMSE		0.45					

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

TABLE C.4: Regression results for the Group scoring treatment when only the first 12 rounds are taken into account.

Self-scoring

	Probit	Linear	OnlyCC	OnlyIR	Only1	Only2	Only3
(Intercept)	-0.12 (0.24)	0.45*** (0.09)	-0.24 (0.14)	0.14 (0.23)	-0.07 (0.24)	-0.01 (0.14)	-0.03 (0.12)
OneCC	0.33*** (0.10)	0.12*** (0.03)	0.33*** (0.10)		0.36*** (0.09)		
TwoCC	-0.08 (0.09)	-0.03 (0.03)	-0.08 (0.09)			0.16* (0.07)	
ThreeCC	0.40*** (0.12)	0.15*** (0.04)	0.41*** (0.12)				0.43*** (0.11)
OneIR	-0.17 (0.21)	-0.06 (0.08)		-0.12 (0.21)	-0.13 (0.21)		
TwoIR	0.01 (0.08)	0.01 (0.03)		0.04 (0.08)		0.05 (0.08)	
ThreeIR	0.08 (0.10)	0.03 (0.04)		0.14 (0.10)			0.10 (0.09)
Learn_Inert	0.36*** (0.07)	0.13*** (0.03)	0.36*** (0.07)	0.37*** (0.07)	0.39*** (0.07)	0.35*** (0.07)	0.33*** (0.07)
tindex	-0.05*** (0.01)	-0.02 (0.01)	-0.05*** (0.01)	-0.06*** (0.01)	-0.06*** (0.01)	-0.06*** (0.01)	-0.05*** (0.01)
AIC	1766.59		1761.84	1788.08	1771.58	1783.82	1771.26
BIC	1813.53		1793.13	1819.37	1797.66	1809.90	1797.34
Log Likelihood	-874.30		-874.92	-888.04	-880.79	-886.91	-880.63
Deviance	1748.59		1749.84	1776.08	1761.58	1773.82	1761.26
Num. obs.	1360	1360	1360	1360	1360	1360	1360
R ²		0.07					
Adj. R ²		0.06					
RMSE		0.48					

***p < 0.001, **p < 0.01, *p < 0.05

TABLE C.5: Regression results for the Self scoring treatment when only the first 12 rounds are taken into account.

Image self-scoring

	Probit	Linear	OnlyCC	OnlyIR	Only1	Only2	Only3
(Intercept)	-0.48 (0.26)	0.32*** (0.09)	-0.18 (0.23)	-0.54* (0.25)	-0.32 (0.25)	-0.53* (0.22)	-0.15 (0.18)
OneCC	-0.22 (0.18)	-0.08 (0.06)	-0.08 (0.16)		-0.01 (0.16)		
TwoCC	0.23 (0.16)	0.08 (0.06)	0.31* (0.15)			0.22 (0.13)	0.38 (0.20)
ThreeCC	0.15 (0.22)	0.05 (0.08)	0.31 (0.21)				
OneIR	0.12 (0.18)	0.04 (0.06)		0.10 (0.17)	0.32 (0.17)		
TwoIR	0.45** (0.15)	0.17** (0.06)		0.49** (0.15)		0.50*** (0.14)	
ThreeIR	0.31 (0.29)	0.11 (0.10)		0.38 (0.28)			0.52 (0.28)
Learn_Inert	0.33* (0.13)	0.12* (0.05)	0.37** (0.13)	0.37** (0.13)	0.44*** (0.13)	0.37** (0.13)	0.37** (0.13)
tindex	-0.01 (0.02)	-0.00 (0.01)	-0.04 (0.02)	-0.01 (0.02)	-0.04 (0.02)	-0.01 (0.02)	-0.04 (0.02)
AIC	585.61		593.47	583.71	597.94	581.26	592.25
BIC	622.59		618.13	608.37	618.49	601.80	612.80
Log Likelihood	-283.80		-290.74	-285.86	-293.97	-285.63	-291.13
Deviance	567.61		581.47	571.71	587.94	571.26	582.25
Num. obs.	450	450	450	450	450	450	450
R ²		0.10					
Adj. R ²		0.08					
RMSE		0.47					

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

TABLE C.6: Regression results for the Image self scoring treatment when only the first 12 rounds are taken into account.

C.4.2 *Fit for all rounds*

In this sections we present the results of the regressions when the entire dataset is taken into consideration. The model with which we fit the data is exactly the same as the one presented in the previous subsection (except, of course, for *tindex* that in this case ranges from 3 to 20).

No scoring

	Probit	Linear	OnlyCC	Only1	Only2	Only3
(Intercept)	-0.24 (0.14)	0.41*** (0.05)	-0.24 (0.14)	-0.19 (0.13)	0.03 (0.12)	0.02 (0.11)
OneCC	0.33*** (0.10)	0.12*** (0.03)	0.33*** (0.10)	0.35*** (0.09)		
TwoCC	-0.08 (0.09)	-0.03 (0.03)	-0.08 (0.09)		0.16* (0.07)	
ThreeCC	0.41*** (0.12)	0.16*** (0.04)	0.41*** (0.12)			0.44*** (0.11)
Learn_Inert	0.36*** (0.07)	0.13*** (0.03)	0.36*** (0.07)	0.39*** (0.07)	0.35*** (0.07)	0.33*** (0.07)
tindex	-0.05*** (0.01)	-0.02*** (0.00)	-0.05*** (0.01)	-0.06*** (0.01)	-0.06*** (0.01)	-0.06*** (0.01)
AIC	1761.84		1761.84	1769.99	1782.18	1770.32
BIC	1793.13		1793.13	1790.85	1803.05	1791.18
Log Likelihood	-874.92		-874.92	-880.99	-887.09	-881.16
Deviance	1749.84		1749.84	1761.99	1774.18	1762.32
Num. obs.	1360	1360	1360	1360	1360	1360
R ²		0.07				
Adj. R ²		0.06				
RMSE		0.48				

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

TABLE C.7: Regression results for the No scoring treatment when all the rounds are taken into account. Due to the lack of information about the scores of players in this treatment, the model considered here doesn't contain the indirect reciprocity dummy variables.

Image scoring

	Probit	Linear	OnlyCC	OnlyIR	Only1	Only2	Only3
(Intercept)	-0.76*** (0.18)	0.22*** (0.06)	-0.12 (0.15)	-0.94*** (0.16)	-0.56** (0.17)	-0.78*** (0.15)	-0.50*** (0.13)
OneCC	-0.23* (0.10)	-0.07* (0.04)	-0.12 (0.10)		-0.10 (0.09)		
TwoCC	-0.02 (0.11)	-0.00 (0.04)	0.12 (0.10)			-0.04 (0.09)	
ThreeCC	0.08 (0.16)	0.02 (0.05)	0.27 (0.15)				0.11 (0.14)
OneIR	0.14 (0.13)	0.04 (0.04)		0.12 (0.12)	0.60*** (0.11)		
TwoIR	0.65*** (0.11)	0.23*** (0.04)		0.63*** (0.10)		0.82*** (0.09)	
ThreeIR	0.42*** (0.12)	0.16*** (0.04)		0.40*** (0.12)			0.76*** (0.11)
Learn_Inert	0.42*** (0.08)	0.14*** (0.03)	0.45*** (0.08)	0.44*** (0.08)	0.45*** (0.08)	0.45*** (0.08)	0.44*** (0.08)
tindex	-0.01 (0.02)	-0.00 (0.01)	-0.04** (0.01)	-0.00 (0.02)	-0.04** (0.01)	-0.01 (0.01)	-0.02 (0.01)
AIC	1275.68		1376.72	1275.54	1350.92	1285.87	1326.07
BIC	1320.38		1406.52	1305.34	1375.75	1310.70	1350.90
Log Likelihood	-628.84		-682.36	-631.77	-670.46	-637.93	-658.03
Deviance	1257.68		1364.72	1263.54	1340.92	1275.87	1316.07
Num. obs.	1060	1060	1060	1060	1060	1060	1060
R ²		0.15					
Adj. R ²		0.14					
RMSE		0.45					

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

TABLE C.8: Regression results for the Image scoring treatment when all the rounds are taken into account.

Group scoring

	Probit	Linear	OnlyCC	OnlyIR	Only1	Only2	Only3
(Intercept)	-0.81*** (0.19)	0.21*** (0.06)	-0.54** (0.17)	-0.79*** (0.18)	-0.77*** (0.19)	-0.55*** (0.16)	-0.30* (0.14)
OneCC	0.09 (0.12)	0.03 (0.04)	0.23* (0.11)		0.13 (0.11)		
TwoCC	-0.07 (0.14)	-0.02 (0.05)	0.12 (0.13)			-0.02 (0.12)	
ThreeCC	-0.23 (0.25)	-0.09 (0.09)	-0.02 (0.24)				-0.01 (0.23)
OneIR	0.26* (0.12)	0.08* (0.04)		0.28* (0.11)	0.38*** (0.11)		
TwoIR	0.49*** (0.14)	0.19*** (0.05)		0.46*** (0.13)		0.61*** (0.13)	
ThreeIR	0.51 (0.40)	0.20 (0.14)		0.42 (0.39)			0.82* (0.39)
Learn_Inert	0.26* (0.10)	0.09* (0.03)	0.30** (0.10)	0.25* (0.10)	0.29** (0.10)	0.25* (0.10)	0.27** (0.10)
tindex	-0.02 (0.02)	-0.01 (0.01)	-0.04* (0.02)	-0.01 (0.02)	-0.02 (0.02)	-0.02 (0.02)	-0.04* (0.02)
AIC	910.91		931.29	906.62	919.01	911.74	932.04
BIC	952.73		959.17	934.50	942.24	934.97	955.27
Log Likelihood	-446.46		-459.65	-447.31	-454.50	-450.87	-461.02
Deviance	892.91		919.29	894.62	909.01	901.74	922.04
Num. obs.	770	770	770	770	770	770	770
R ²		0.07					
Adj. R ²		0.06					
RMSE		0.45					

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

TABLE C.9: Regression results for the Group scoring treatment when all the rounds are taken into account.

Self-scoring

	Probit	Linear	OnlyCC	OnlyIR	Only1	Only2	Only3
(Intercept)	-0.12 (0.24)	0.45*** (0.09)	-0.24 (0.14)	0.14 (0.23)	-0.07 (0.24)	-0.01 (0.14)	-0.03 (0.12)
OneCC	0.33*** (0.10)	0.12*** (0.03)	0.33*** (0.10)		0.36*** (0.09)		
TwoCC	-0.08 (0.09)	-0.03 (0.03)	-0.08 (0.09)			0.16* (0.07)	
ThreeCC	0.40*** (0.12)	0.15*** (0.04)	0.41*** (0.12)				0.43*** (0.11)
OneIR	-0.17 (0.21)	-0.06 (0.08)		-0.12 (0.21)	-0.13 (0.21)		
TwoIR	0.01 (0.08)	0.01 (0.03)		0.04 (0.08)		0.05 (0.08)	
ThreeIR	0.08 (0.10)	0.03 (0.04)		0.14 (0.10)			0.10 (0.09)
Learn_Inert	0.36*** (0.07)	0.13*** (0.03)	0.36*** (0.07)	0.37*** (0.07)	0.39*** (0.07)	0.35*** (0.07)	0.33*** (0.07)
tindex	-0.05*** (0.01)	-0.02 (0.01)	-0.05*** (0.01)	-0.06*** (0.01)	-0.06*** (0.01)	-0.06*** (0.01)	-0.05*** (0.01)
AIC	1766.59		1761.84	1788.08	1771.58	1783.82	1771.26
BIC	1813.53		1793.13	1819.37	1797.66	1809.90	1797.34
Log Likelihood	-874.30		-874.92	-888.04	-880.79	-886.91	-880.63
Deviance	1748.59		1749.84	1776.08	1761.58	1773.82	1761.26
Num. obs.	1360	1360	1360	1360	1360	1360	1360
R ²		0.07					
Adj. R ²		0.06					
RMSE		0.48					

***p < 0.001, **p < 0.01, *p < 0.05

TABLE C.10: Regression results for the Self scoring treatment when all the rounds are taken into account.

Image self-scoring

	Probit	Linear	OnlyCC	OnlyIR	Only1	Only2	Only3
(Intercept)	-0.48 (0.26)	0.32*** (0.09)	-0.18 (0.23)	-0.54* (0.25)	-0.32 (0.25)	-0.53* (0.22)	-0.15 (0.18)
OneCC	-0.22 (0.18)	-0.08 (0.06)	-0.08 (0.16)		-0.01 (0.16)		
TwoCC	0.23 (0.16)	0.08 (0.06)	0.31* (0.15)			0.22 (0.13)	
ThreeCC	0.15 (0.22)	0.05 (0.08)	0.31 (0.21)				0.38 (0.20)
OneIR	0.12 (0.18)	0.04 (0.06)		0.10 (0.17)	0.32 (0.17)		
TwoIR	0.45** (0.15)	0.17** (0.06)		0.49** (0.15)		0.50*** (0.14)	
ThreeIR	0.31 (0.29)	0.11 (0.10)		0.38 (0.28)			0.52 (0.28)
Learn_Inert	0.33* (0.13)	0.12* (0.05)	0.37** (0.13)	0.37** (0.13)	0.44*** (0.13)	0.37** (0.13)	0.37** (0.13)
tindex	-0.01 (0.02)	-0.00 (0.01)	-0.04 (0.02)	-0.01 (0.02)	-0.04 (0.02)	-0.01 (0.02)	-0.04 (0.02)
AIC	585.61		593.47	583.71	597.94	581.26	592.25
BIC	622.59		618.13	608.37	618.49	601.80	612.80
Log Likelihood	-283.80		-290.74	-285.86	-293.97	-285.63	-291.13
Deviance	567.61		581.47	571.71	587.94	571.26	582.25
Num. obs.	450	450	450	450	450	450	450
R ²		0.10					
Adj. R ²		0.08					
RMSE		0.47					

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

TABLE C.11: Regression results for the Image Self scoring treatment when all the rounds are taken into account.

C.4.3 Marginal effects

In the following subsection, we plot the marginal effects of each independent variable over the dependent variable, i.e. the choice to cooperate or defect in the following round, as obtained from the probit regression⁸.

For each treatment, we compute the average of the sample marginal effects, as suggested in Fernihough [193].

The results are shown below; the error bars indicate the confidence interval at 2σ .

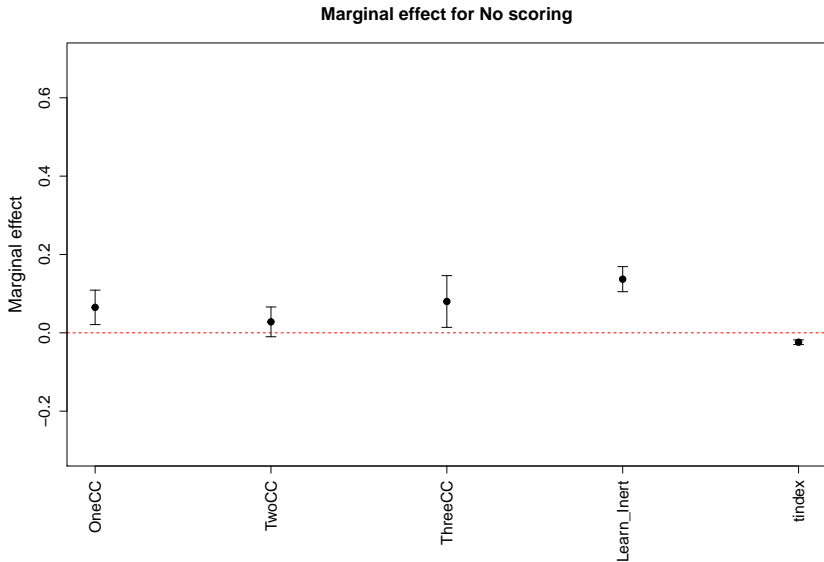


FIGURE C.18: Marginal effects of all the independent variables for the No scoring treatment. Due to the lack of information about the scores of players in this treatment, the model considered here doesn't contain the indirect reciprocity dummy variables.

⁸ The marginal effect m of an independent variable x on a dependent variable y can be interpreted as the value such that a unit increase in x increases y by m units.

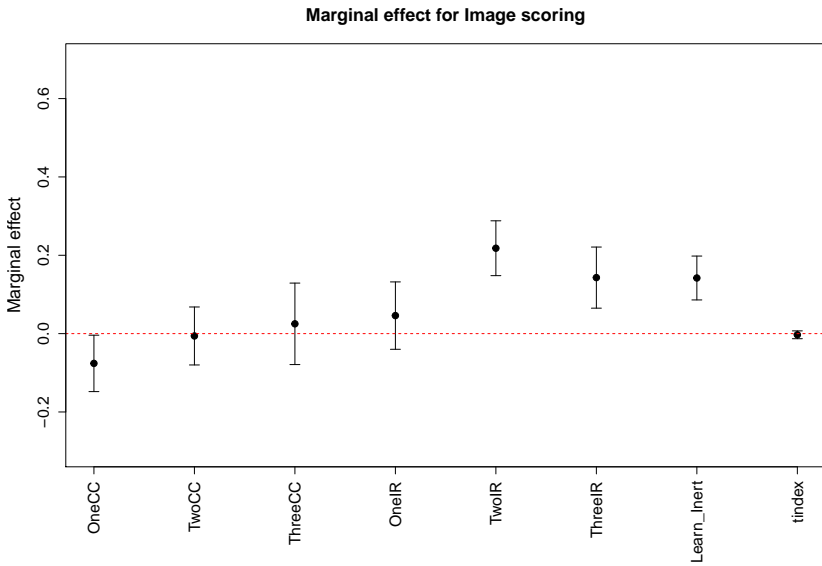


FIGURE C.19: Marginal effects of all the independent variables for the Image scoring treatment.

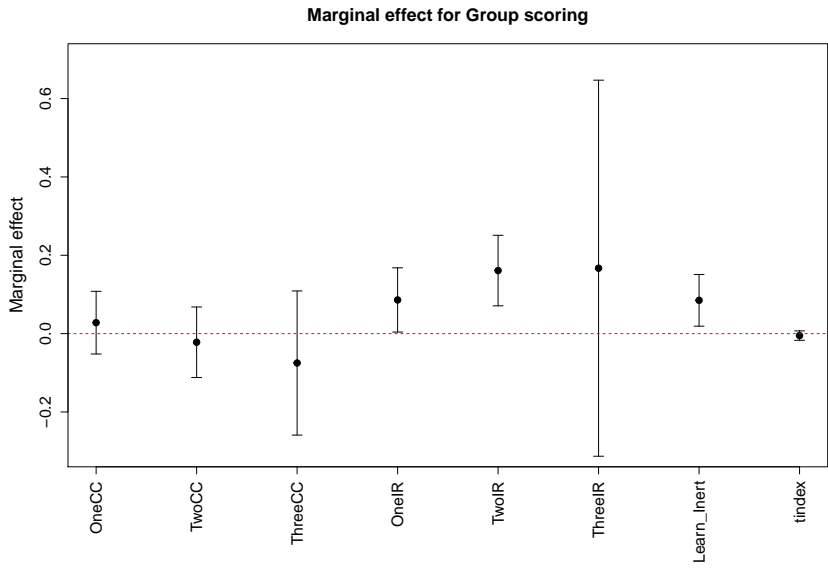


FIGURE C.20: Marginal effects of all the independent variables for the Group scoring treatment.

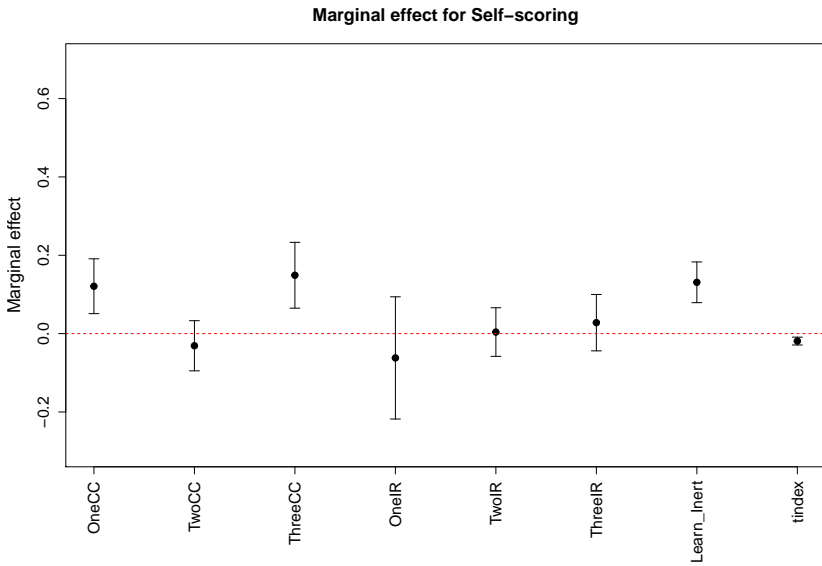


FIGURE C.21: Marginal effects of all the independent variables for the Self scoring treatment.

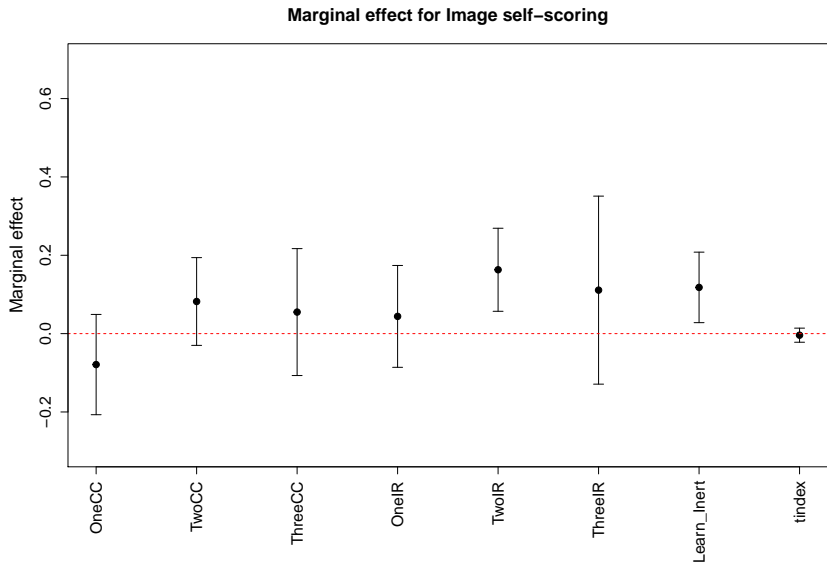


FIGURE C.22: Marginal effects of all the independent variables for the Image self scoring treatment.

C.4.4 *Order Effects*

Throughout both this and the main manuscript, we have assumed that the order in which treatments are played have no influence over the decision of the players. This assumption was supported by qualitative observation of our data and the well known restart effect (see e.g. Andreoni [21]).

Here we very briefly check whether this assumption holds by performing a very simple linear regression with the order of the treatment as the only independent variable and the player's decision as dependent variable:

Order linear regression	
(Intercept)	0.41*** (0.01)
Order	-0.03*** (0.01)
R ²	0.00
Adj. R ²	0.00
Num. obs.	13248
RMSE	0.47

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

TABLE C.12: Results of a linear regression assuming the order in which treatments are played as the only independent variable.

As we can observe from the above table, the order in which the treatments are played results to be significant. However, the marginal effect of it seems to be quite small⁹.

Hence, we conclude that, while the order is not completely irrelevant, its effect on the overall decision of the players are small enough that our approximation remains reasonable.

⁹ In a linear regression, the coefficient of the regression is a good approximation of the marginal effect of an independent variable.

C.5 SCREENSHOTS

In the following we show screenshots of various stages of the experiment. These are exactly what the participants saw on their screens.

Payoff calculator

Period: TRUL out of 1 Remaining time (sec): 103

You now have some time to familiarize yourself with how the payoff works. Please note that none of the choices you take at this stage will influence your final earnings.

Choose an action for you and one for each of your groupmates and you will be shown the payoff that corresponds to these actions.

Choose the actions.

You private account group account

Groupmate 1 private account group account

Groupmate 3 private account group account

Groupmate 4 private account group account

Starting with an endowment of 1.0, you earn **2.0** EU.

Leave payoff calculator

FIGURE C.23: This is a screenshot of the payoff calculator. In every session, before the Initial Phase, players were given time to understand the game by trying different combinations of theirs and their teammates actions and observe the payoff outcome.

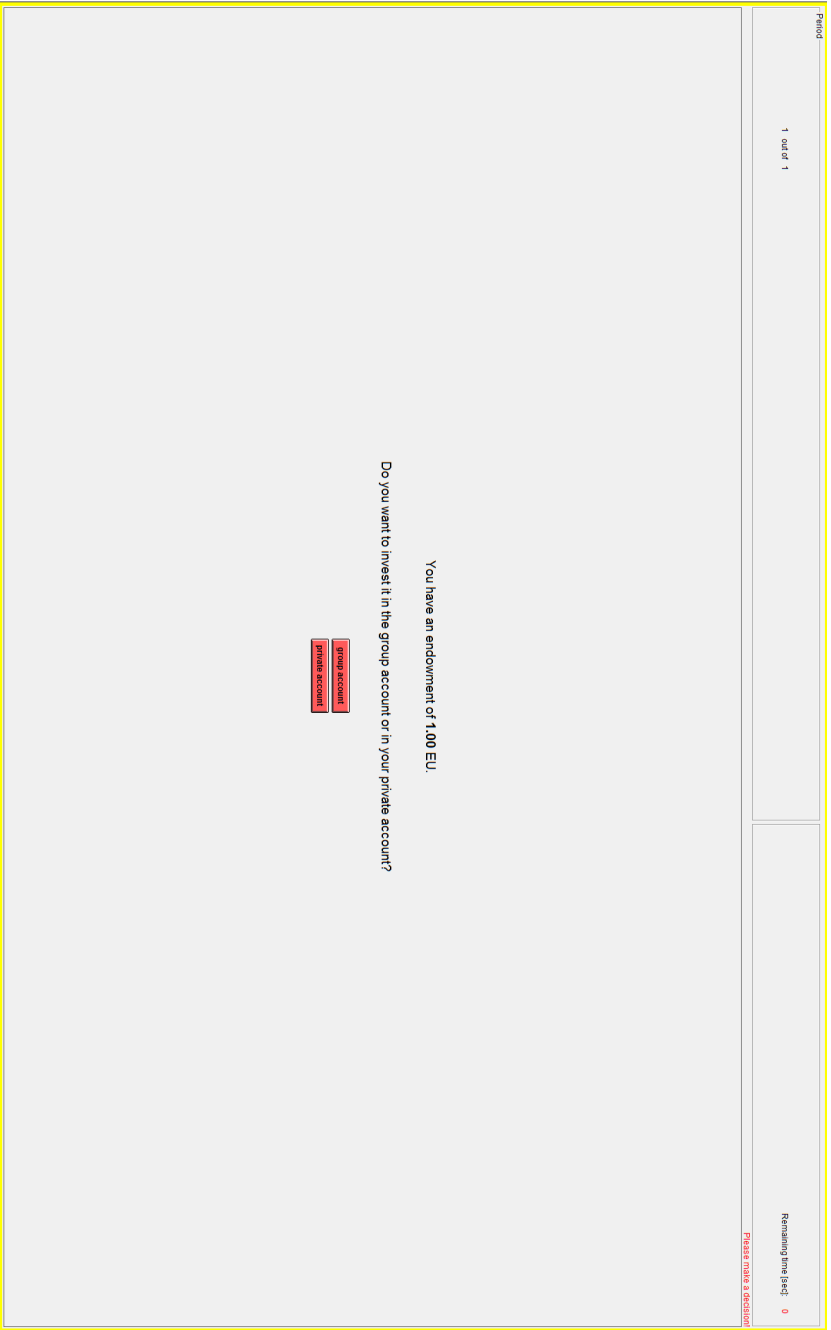
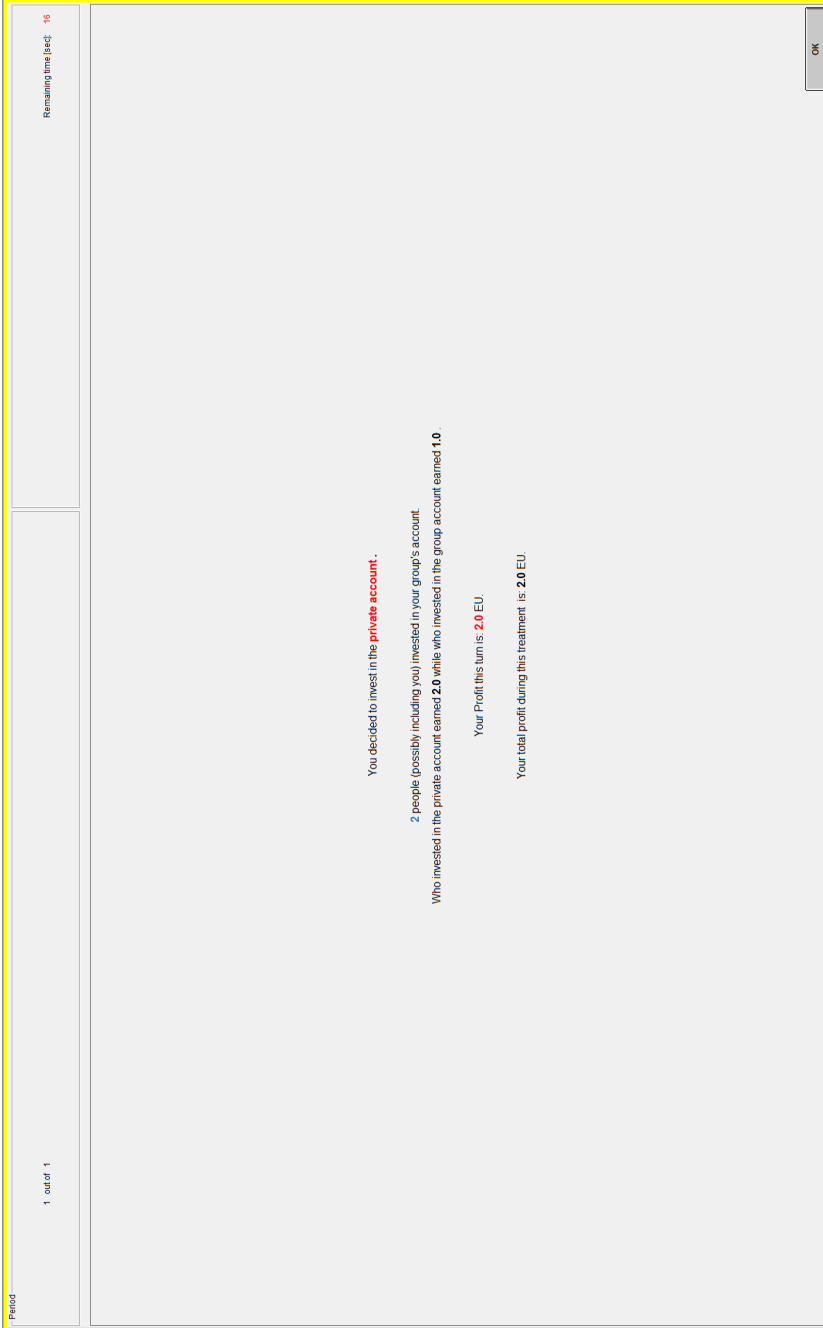


FIGURE C.24: This is the decision screen shown to the players during the initial phase. As shown by the screenshot, no information regarding the teammates is provided to the players.



Feedback

FIGURE C.25: This is the feedback screen as shown in all treatments other than Self Scoring. It shows how many players defected or cooperated in the player's group and his/her resulting payoff for the round.

Image scoring

Periods
2 out of 2

Remaining time (sec) 25

In the last round:	
Earnings:	Scores:
People who invested in the group account in this group: 3:	
1.5	1.00
1.5	1.00
1.5	1.00
2.5	0.00
People who invested in the group account in this group: 1:	
1.5	0.00
0.5	1.00
1.5	0.00
1.5	0.00
People who invested in the group account in this group: 4:	
2.0	1.00
2.0	1.00
2.0	1.00
2.0	1.00
People who invested in the group account in this group: 3:	
1.5	1.00
1.5	1.00
1.5	1.00
2.5	0.00

This is you
These are the others in your group this round.

You have an endowment of 1.00 EU.

Do you want to invest it in the group account or in your private account?

FIGURE C.26: This is the decision screen shown to the players in the Image Scoring treatment (from the second round on). On the left it shows how players were grouped in the previous round and what where the individual decisions. The green line highlights the player and the yellow lines the player's teammates in the current round.

Group scoring

Finished

2 out of 2

Remaining time 8 Sec 25

In the last round:

People who invested in the group account in this group: 3.

This player has score	0.75
This player has score	0.75
This player has score	0.75
This player has score	0.75

People who invested in the group account in this group: 1.

This player has score	0.25
This player has score	0.25
This player has score	0.25
This player has score	0.25
This player has score	0.25
This player has score	0.25

People who invested in the group account in this group: 1.

This player has score	0.25
This player has score	0.25
This player has score	0.25
This player has score	0.25

People who invested in the group account in this group: 2.

This player has score	0.50
This player has score	0.50
This player has score	0.50
This player has score	0.50

This is you.
These are the players in your group this rounds.

You have an endowment of 1.00 EU.

Do you want to invest it in the group account or in your private account?

group account

private account

FIGURE C.27: This is the decision screen shown to the players in the Group Scoring treatment (from the second round on). On the left it shows how players were grouped in the previous round and the group score of each group. The green line highlights the player and the yellow lines the player's teammates in the current round.

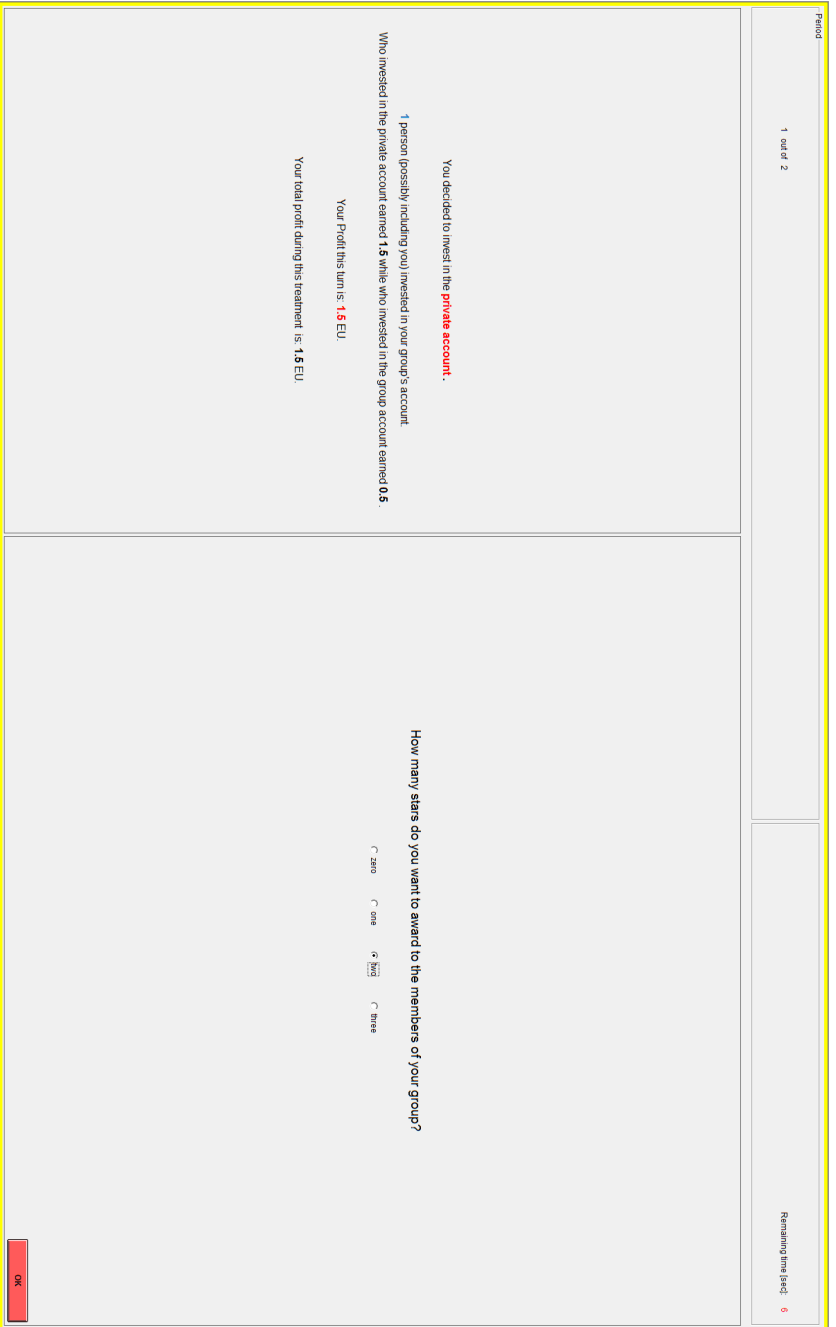


FIGURE C.28: This is the feedback screen shown to the players in the Self Scoring treatment. In addition to receive the same feedback as in all the other treatments, players are asked to rate their group on a scale from zero to three.

Image Self scoring

Period
2 out of 2
Remaining time (sec): 27

In the last round:

Earnings:	Scores:
People who invested in the group account in this group: 2.	
2.0	0.44
1.0	0.56
1.0	0.56
2.0	0.44
People who invested in the group account in this group: 2.	
1.0	0.56
1.0	0.56
2.0	0.44
2.0	0.44
People who invested in the group account in this group: 2.	
1.0	0.56
1.0	0.56
2.0	0.44
2.0	0.44
People who invested in the group account in this group: 1.	
1.5	0.22
1.5	0.22
0.5	0.33
1.5	0.22

You have an endowment of 1.00 EU.

Do you want to invest it in the group account or in your private account?

group account
private account

This is you.
These are the players in your group this round.

FIGURE C.29: This is the decision screen shown to the players in the Image Self Scoring treatment (from the second round on). On the left it shows how players where grouped in the previous round and each player's (image-self) score. The green line highlights the player and the yellow lines the player's teammates in the current round.

Image self-score calculator

Period

TRIAL out of 1

Remaining time (sec) 1

You now have some time to familiarize yourself with how the score system works. Please note that none of the choices you take at this stage will influence your final earnings.

Choose an action for you and one for each of your groupmates and you will be shown the scores that correspond to these actions.

Choose the actions.

You private account group account

Groupmate 1 private account group account

Groupmate 3 private account group account

Groupmate 4 private account group account

With these choice of actions the score of each player would be:

You	Groupmate 1	Groupmate 2	Groupmate 3
4/9 = 0.44	5/9 = 0.56	4/9 = 0.44	5/9 = 0.56

[Leave score calculator](#)

Figure C.30: This is a screenshot of the Image Self score calculator. Before the Image Self score treatment, players were given time to understand the scoring mechanism by trying different combinations of theirs and their teammates actions and observe the resulting score.

BIBLIOGRAPHY

1. Roger, B. M. Game theory: Analysis of conflict. *The President and Fellows of Harvard College, USA* (1991).
2. Neumann, J. v. Zur theorie der gesellschaftsspiele. *Mathematische Annalen* **100**, 295 (1928).
3. Von Neumann, J. & Morgenstern, O. Theory of Games and Economic Behavior. *Princeton University Press*, 625 (1944).
4. Gibbons, R. *A primer in game theory* (Harvester Wheatsheaf, 1992).
5. Arrow, K. J. *Economic theory and the hypothesis of rationality in Utility and Probability* (eds Eatwell, J., Milgate, M. & Newman, P.) 25 (Springer, 1990).
6. Nash, J. Equilibrium points in n-person games. *Proceedings of the National Academy of Sciences* **36**, 48 (1950).
7. Ostrom, E. Coping with tragedies of the commons. *Annual Review of Political Science* **2**, 493 (1999).
8. Hardin, G. The tragedy of the commons. *Science* **1243** (1968).
9. Archetti, M. How to Analyze Models of Nonlinear Public Goods. *Games* **9** (2018).
10. Buchanan, J. M. An economic theory of clubs. *Economica* **32**, 1 (1965).
11. Paúly, M. V. Clubs, Commonality, and the Core: An Integration of Game Theory and the Theory of Public Goods. *Economica* **34**, 314 (1967).
12. McConnell, C. R., Brue, S. L. & Flynn, S. M. *Economics: Principles, problems, and policies* (Boston McGraw-Hill/Irwin, 2009).
13. Cox, J. C. & Sadiraj, V. On modeling voluntary contributions to public goods. *Public Finance Review* **35**, 311 (2007).
14. Marwell, G. & Ames, R. E. Experiments on the provision of public goods. I. Resources, interest, group size, and the free-rider problem. *American Journal of Sociology* **84**, 1335 (1979).
15. Isaac, R. M. & Walker, J. M. Group Size Effects in Public Goods Provision: The Voluntary Contributions Mechanism. *The Quarterly Journal of Economics* **103**, 179 (1988).

16. Palfrey, T. R. & Prisbrey, J. E. Anomalous behavior in public goods experiments: How much and why? *The American Economic Review*, 829 (1997).
17. Andreoni, J. & Miller, J. Giving according to GARP: An experimental test of the consistency of preferences for altruism. *Econometrica* 70, 737 (2002).
18. Diekmann, A. *Volunteer's Dilemma. A Social Trap without a Dominant Strategy and some Empirical Results in Paradoxical Effects of Social Behavior: Essays in Honor of Anatol Rapoport* (eds Diekmann, A. & Mitter, P.) 187 (Physica-Verlag HD, Heidelberg, DE, 1986).
19. Ledyard, J. O. *Public goods: A survey of Experimental Research in The Handbook of Experimental Economics* 111 (Princeton University Press, 1995).
20. Chaudhuri, A. Sustaining cooperation in laboratory public goods experiments: A selective survey of the literature. *Experimental Economics* 14, 47 (2011).
21. Andreoni, J. Why free ride? Strategies and learning in public goods experiments. *Journal of Public Economics* 37, 291 (1988).
22. Fudenberg, D. & Tirole, J. *Game theory* 12, 80 (MIT Press, Cambridge, MA, 1991).
23. Mas-Colell, A., Whinston, M. D., Green, J. R., et al. *Microeconomic theory* (Oxford University Press, 1995).
24. Feigenbaum, J. & Shenker, S. *Distributed Algorithmic Mechanism Design: Recent Results and Future Directions in Proceedings of the 6th International Workshop on Discrete Algorithms and Methods for Mobile Computing and Communications* (ACM, Atlanta, Georgia, USA, 2002), 1.
25. Erdman, A. G., Sandor, G. N. & Kota, S. *Mechanism design: Analysis and synthesis. Vol. 1* (Prentice-Hall, 1997).
26. Abdulkadirouglu, A. & Sönmez, T. School choice: A mechanism design approach. *American Economic Review* 93, 729 (2003).
27. Marden, J. R. & Shamma, J. S. *Game theory and distributed control in Handbook of game theory with economic applications* 861 (Elsevier, 2015).
28. Williams, A. The optimal provision of public goods in a system of local government. *Journal of Political Economy* 74, 18 (1966).
29. Andreoni, J. & Bergstrom, T. Do government subsidies increase the private supply of public goods? *Public Choice* 88, 295 (1996).

30. Ostrom, E. *Governing the commons: The Evolution of Institutions for Collective Action* (Cambridge University Press, Cambridge, U.K., 1990).
31. Johnson, D. L. Nomadism and desertification in Africa and the Middle East. *GeoJournal* **31**, 51 (1993).
32. Rege, M. Social norms and private provision of public goods. *Journal of Public Economic Theory* **6**, 65 (2004).
33. Traulsen, A., Röhl, T. & Milinski, M. An economic experiment reveals that humans prefer pool punishment to maintain the commons. *Proceedings of the Royal Society of London B: Biological Sciences* (2012).
34. Gürerk, Ö., Irlenbusch, B. & Rockenbach, B. The competitive advantage of sanctioning institutions. *Science* **312**, 108 (2006).
35. Helbing, D., Szolnoki, A., Perc, M. & Szabó, G. Punish, but not too hard: How costly punishment spreads in the spatial public goods game. *New Journal of Physics* **12**, 083005 (2010).
36. McCusker, C. & Carnevale, P. J. Framing in resource dilemmas: Loss aversion and the moderating effects of sanctions. *Organizational Behavior and Human Decision Processes* **61**, 190 (1995).
37. Sigmund, K., De Silva, H., Traulsen, A. & Hauert, C. Social learning promotes institutions for governing the commons. *Nature* **466**, 861 (2010).
38. Gunnthorsdottir, A., Vragov, R., Seifert, S. & McCabe, K. Near-efficient equilibria in contribution-based competitive grouping. *Journal of Public Economics* **94**, 987 (2010).
39. Gunnthorsdottir, A., Vragov, R. & Shen, J. Tacit coordination in contribution-based grouping with two endowment levels. *Research in Experimental Economics* **13**, 13 (2010).
40. Nax, H. H., Baliotti, S., Murphy, R. O. & Helbing, D. Adding noise to the institution: An experimental welfare investigation of the contribution-based grouping mechanism. *Social Choice and Welfare* **50**, 213 (2018).
41. Rabanal, J. P. & Rabanal, O. A. Efficient Investment via Assortative Matching: A laboratory experiment. *mimeo* (2014).
42. Andreoni, J. Cooperation in public-goods experiments: kindness or confusion? *The American Economic Review* **85**, 891 (1995).

43. Fischbacher, U. & Gächter, S. Social preferences, beliefs, and the dynamics of free riding in public good experiments. *American Economic Review* **100**, 541 (2010).
44. Frey, B. S. & Meier, S. Social comparisons and pro-social behavior: Testing "conditional cooperation" in a field experiment. *The American Economic Review* **94**, 1717 (2004).
45. Kocher, M. G., Cherry, T., Kroll, S., Netzer, R. J. & Sutter, M. Conditional cooperation on three continents. *Economics Letters* **101**, 175 (2008).
46. Ockenfels, A. *Fairneß, Reziprozität und Eigennutz: Ökonomische Theorie und experimentelle Evidenz* (Mohr Siebeck, 1999).
47. Fischbacher, U., Gächter, S. & Fehr, E. Are people conditionally cooperative? Evidence from a public goods experiment. *Economics Letters* **71**, 397 (2001).
48. Becker, L. C. *Reciprocity* (Routledge, 2014).
49. Nowak, M. A. & Sigmund, K. Evolution of indirect reciprocity. *Nature* **437**, 1291 (2005).
50. Binmore, K. G. *Game theory and the social contract* (MIT Press, Cambridge, MA, 1984).
51. Gouldner, A. W. The norm of reciprocity: A preliminary statement. *American Sociological Review* **25**, 161 (1960).
52. Binford, L. R. Human ancestors: Changing views of their behavior. *Journal of Anthropological Archaeology* **4**, 292 (1985).
53. Fehr, E. & Gächter, S. Fairness and Retaliation: The Economics of Reciprocity. *Journal of Economic Perspectives* **14**, 159 (2000).
54. Fudenberg, D. & Maskin, E. The Folk Theorem in Repeated Games with Discounting or with Incomplete Information. *Econometrica* **54**, 533 (1986).
55. Binmore, K. *Playing for Real Coursepack Edition: A Text on Game Theory* (Oxford University Press, Oxford, U.K., 2012).
56. McGillivray, F. & Smith, A. Trust and cooperation through agent-specific punishments. *International Organization* **54**, 809 (2000).
57. Rosenthal, R. W. Sequences of Games with Varying Opponents. *Econometrica* **47**, 1353 (1979).
58. Kandori, M. Social Norms and Community Enforcement. *The Review of Economic Studies* **59**, 63 (1992).

59. Nowak, M. & Highfield, R. *Supercooperators: Altruism, evolution, and why we need each other to succeed* (Simon and Schuster, New York, NY, 2011).
60. Milinski, M. Reputation, a universal currency for human social interactions. *Philosophical Transactions of the Royal Society B: Biological Sciences* **371**, 20150100 (2016).
61. Nowak, M. A. & Sigmund, K. Evolution of indirect reciprocity by image scoring. *Nature* **393**, 573 (1998).
62. Nowak, M. A. & Sigmund, K. The dynamics of indirect reciprocity. *Journal of Theoretical Biology* **194**, 561 (1998).
63. Sugden, R. *The economics of rights, co-operation and welfare* (Springer, 1986).
64. Leimar, O. & Hammerstein, P. Evolution of cooperation through indirect reciprocity. *Proceedings of Biological Sciences. The Royal Society* **268**, 745 (2001).
65. Ohtsuki, H. & Iwasa, Y. The leading eight: Social norms that can maintain cooperation by indirect reciprocity. *Journal of Theoretical Biology* **239**, 435 (2006).
66. Wedekind, C. & Milinski, M. Cooperation through image scoring in humans. *Science* **288**, 850 (2000).
67. Wedekind, C. & Braithwaite, V. A. The long-term benefits of human generosity in indirect reciprocity. *Current Biology* **12**, 1012 (2002).
68. Seinen, I. & Schram, A. Social status and group norms: Indirect reciprocity in a repeated helping experiment. *European Economic Review* **50**, 581 (2006).
69. Bolton, G. E., Katok, E. & Ockenfels, A. Cooperation among strangers with limited information about reputation. *Journal of Public Economics* **89**, 1457 (2005).
70. Milinski, M., Semmann, D., Bakker, T. C. & Krambeck, H.-J. Cooperation through indirect reciprocity: Image scoring or standing strategy? *Proceedings of the Royal Society of London B: Biological Sciences* **268**, 2495 (2001).
71. Nax, H. H., Perc, M., Szolnoki, A. & Helbing, D. Stability of cooperation under image scoring in group interactions. *Scientific Reports* **5** (2015).

72. Björnerstedt, J. & Weibull, J. W. *Nash equilibrium and evolution by imitation* tech. rep. (IUI Working Paper, 1994).
73. Smith, J. M. *Evolution and the Theory of Games* (Cambridge University Press, Cambridge, UK, 1982).
74. Foster, D. P. & Vohra, R. V. Calibrated learning and correlated equilibrium. *Games and Economic Behavior* **21**, 40 (1997).
75. Young, H. P. Learning by trial and error. *Games and Economic Behavior* **65**, 626 (2009).
76. Nash, J. Non-cooperative games. *Annals of Mathematics* **54**, 286 (1951).
77. Young, H. P. The evolution of conventions. *Econometrica* **61**, 57 (1993).
78. Kandori, M., Mailath, G. J. & Rob, R. Learning, mutation, and long run equilibria in games. *Econometrica*, 29 (1993).
79. Fudenberg, D. & Levine, D. Learning in games. *European Economic Review* **42**, 631 (1998).
80. Young, H. & Zamir, S. *Handbook of Game Theory* (Elsevier Science, 2014).
81. Blume, L. E. The statistical mechanics of strategic interaction. *Games and Economic Behavior* **5**, 387 (1993).
82. McKelvey, R. D. & Palfrey, T. R. Quantal Response Equilibria for Normal Form Games. *Games and Economic Behavior* **10**, 6 (1995).
83. McKelvey, R. D. & Palfrey, T. R. Quantal response equilibria for extensive form games. *Experimental Economics* **1**, 9 (1998).
84. Gigerenzer, G. & Selten, R. *Bounded rationality: The adaptive toolbox* (MIT Press, Cambridge, MA, 2002).
85. Lahkar, R. & Riedel, F. The logit dynamic for games with continuous strategy sets. *Games and Economic Behavior* **91**, 268 (2015).
86. Cherry, T. L., Kroll, S. & Shogren, J. F. The impact of endowment heterogeneity and origin on public good contributions: Evidence from the lab. *Journal of Economic Behavior and Organization* **57**, 357 (2005).
87. Reuben, E. & Riedl, A. Enforcement of contribution norms in public good games with heterogeneous populations. *Games and Economic Behavior* **77**, 122 (2013).
88. Przepiorka, W. & Diekmann, A. Individual heterogeneity and costly punishment: A volunteer's dilemma. *Proceedings of the Royal Society of London B: Biological Sciences* **280**, 20130247 (2013).

89. Bennati, S. & Pournaras, E. Privacy-enhancing aggregation of Internet of Things data via sensors grouping. *Sustainable Cities and Society* **39**, 387 (2018).
90. Olson, M. *The Logic of Collective Action: Public Goods and the Theory of Groups* (Harvard University Press, Cambridge, MA, 1965).
91. Isaac, R. M., McCue, K. F. & Plott, C. R. Public goods provision in an experimental environment. *Journal of Public Economics* **26**, 51 (1985).
92. Schlager, E. & Ostrom, E. Property-Rights Regimes and Natural Resources: A Conceptual Analysis. *Land Economics* **68**, 249 (1992).
93. Ledyard, J. O. Public Goods: A Survey of Experimental Research. in J. H. Kagel and A. E. Roth (Eds.), *Handbook of experimental economics* **37**, 111 (1995).
94. Nax, H. H., Murphy, R. O. & Helbing, D. *Stability and welfare of 'merit-based' group-matching mechanisms in voluntary contribution games* Risk Center working paper, ETH Zurich. 2014.
95. Fehr, E. & Gächter, S. Cooperation and Punishment in Public Goods Experiments. *American Economic Review* **90**, 980 (2000).
96. Perc, M. Does strong heterogeneity promote cooperation by group interactions? *New Journal of Physics* **13**, 123027 (2011).
97. Harsanyi, J. C. & Selten, R. *A General Theory of Equilibrium Selection in Games* (MIT Press, Cambridge, MA, 1988).
98. Harsanyi, J. C. A New Theory of Equilibrium Selection for Games with Complete Information. *Games and Economic Behavior* **8**, 91 (1995).
99. Abreu, D., Pearce, D. & Stacchetti, E. Toward a theory of discounted repeated games with imperfect monitoring. *Econometrica*, 1041 (1990).
100. Fudenberg, D., Levine, D. & Maskin, E. The folk theorem with imperfect public information. *Econometrica*, 997 (1994).
101. Wu, J. & Axelrod, R. How to cope with noise in the iterated prisoner's dilemma. *Journal of Conflict Resolution* **39**, 183 (1995).
102. Alchian, A. A. & Demsetz, H. Production, information costs, and economic organization. *American Economic Review* **62**, 777 (1972).
103. Andreoni, J. An experimental test of the public goods crowding-out hypothesis. *American Economic Review* **83**, 1317 (1993).
104. Palfrey, T. R. & Prisbrey, J. E. Altruism, reputation and noise in linear public goods experiments. *Journal of Public Economics* **61**, 409 (1996).

105. Palfrey, T. R. & Prisbrey, J. E. Anomalous behavior in public goods experiments: How much and why? *American Economic Review* **87**, 829 (1997).
106. Goeree, J. K., Holt, C. A. & Laury, S. K. Private costs and public benefits: Unraveling the effects of altruism and noisy behavior. *Journal of Public Economics* **83**, 255 (2002).
107. Ferraro, P. J. & Vossler, C. A. The source and significance of confusion in public goods experiments. *The B.E. Journal in Economic Analysis and Policy* **10**, 53 (2010).
108. Bayer, R.-C., Renner, E. & Sausgruber, R. Confusion and learning in the voluntary contributions game. *Experimental Economics* **16**, 478 (2013).
109. Nax, H. H., Burton-Chellew, M. N., West, S. A. & Young, H. P. Learning in a black box. *Journal of Economic Behavior & Organization* **127**, 1 (2016).
110. Burton-Chellew, M. N., Nax, H. H. & West, S. A. Payoff-based learning explains the decline in cooperation in public goods games. *Proceedings of the Royal Society of London B: Biological Sciences* **282**, 20142678 (2015).
111. Gunnthorsdottir, A. & Thorsteinsson, P. Tacit Coordination and Equilibrium Selection in a Merit-Based Grouping Mechanism: A Cross-Cultural Validation Study. *Department of Economics WP* (2010).
112. Rud, O. & Rabanal, J. P. Efficient Investment via Assortative Matching (2015).
113. Ones, U. & Putterman, L. The ecology of collective action: A public goods and sanctions experiment with controlled group formation. *Journal of Economic Behavior and Organization* **62**, 495 (2007).
114. Harsanyi, J. Cardinal Utility in Welfare Economics and in the Theory of Risk-Taking. *Journal of Political Economy* **61**, 434 (1953).
115. Binmore, K. *Natural Justice* (Oxford University Press, Oxford, UK, 2005).
116. Harsanyi, J. C. & Selten, R. *A General Theory of Equilibrium Selection in Games* (MIT Press, Cambridge, MA, 1988).
117. Atkinson, A. B. On the measurement of inequality. *Journal of Economic Theory* **2**, 244 (1970).

118. Jones-Lee, M. W. & Loomes, G. Discounting and Safety. *Oxford Economic Papers. New Series* **47**, 501 (1995).
119. Dorfman, R. A formula for the Gini coefficient. *The Review of Economics and Statistics*, 146 (1979).
120. Gini, C. Concentration and dependency ratios. *Rivista di Politica Economica* **87**, 769 (1997).
121. Mäs, M. & Nax, H. H. A behavioral study of “noise” in coordination games. *Journal of Economic Theory* **162**, 195 (2016).
122. Cinyabuguma, M., Page, T. & Putterman, L. Cooperation under the threat of expulsion in a public goods experiment. *Journal of Public Economics* **89**, 1421 (2005).
123. Charness, G. B. & Yang, C.-L. Endogenous Group Formation and Public Goods Provision: Exclusion, Exit, Mergers, and Redemption. *University of California at Santa Barbara, Economics WP* (2008).
124. Ehrhart, K. & Keser, C. Mobility and cooperation: On the run. *CIRANO Working Paper* **99** (1999).
125. Ahn, T., Isaac, R. M. & Salmon, T. C. Endogenous group formation. *Journal of Public Economic Theory* **10**, 171 (2008).
126. Coricelli, G., Fehr, D. & Fellner, G. Partner Selection in Public Goods Experiments. *Economics Series* **151** (2004).
127. Page, T., Putterman, L. & Unel, B. Voluntary association in public goods experiments: Reciprocity, Mimicry and Efficiency. *The Economic Journal* **115**, 1032 (2005).
128. Brekke, K., Nyborg, K. & Rege, M. The fear of exclusion: Individual effort when group formation is endogenous. *Scandinavian Journal of Economics* **109**, 531 (2007).
129. Brekke, K., Hauge, K., Lind, J. T. & Nyborg, K. Playing with the good guys. A public good game with endogenous group formation. *Journal of Public Economics* **95**, 1111 (2011).
130. Simon, H. A. A mechanism for social selection and successful altruism. *Science* **250**, 1665 (1990).
131. Bowles, S. & Gintis, H. *A cooperative species—human reciprocity and its evolution* (Princeton University Press, Princeton, NJ, 2011).
132. Fehr, E. & Camerer, C. Social neuroeconomics: the neural circuitry of social preferences. *Trends in Cognitive Sciences* **11**, 419 (2007).

133. Allchin, D. The Evolution of Morality. *Evolution: Education and Outreach* **2**, 590 (2009).
134. Hamilton, W. D. The Genetical Evolution of Social Behaviour I. *Journal of Theoretical Biology* **7**, 1 (1964).
135. Hamilton, W. D. The Genetical Evolution of Social Behaviour II. *Journal of Theoretical Biology* **7**, 17 (1964).
136. Nowak, M. A. Five rules for the evolution of cooperation. *Science* **314**, 1560 (2006).
137. Alger, I. & Weibull, J. Homo Moralis - Preference Evolution Under Incomplete Information and Assortative Matching. *Econometrica* **81**, 2269 (2013).
138. Grund, T., Waloszek, C. & Helbing, D. How Natural Selection Can Create Both Self- and Other-Regarding Preferences, and Networked Minds. *Scientific Reports* **3**, 1480 (2013).
139. Newton, J. The preferences of Homo Moralis are unstable under evolving assortativity. *International Journal of Game Theory* **46**, 583 (2017).
140. Nax, H. H. & Rigos, A. Assortativity evolving from social dilemmas. *Journal of Theoretical Biology* **395**, 194 (2016).
141. Newton, J. Shared intentions: The evolution of collaboration. *Games and Economic Behavior* **104**, 517 (2017).
142. Cabral, L. M. B. Asymmetric equilibria in symmetric games with many players. *Economic Letters* **27**, 205 (1988).
143. Gardiner, C. *Handbook of stochastic methods for physics, chemistry, and the natural sciences* (Springer, Berlin, Germany, 1994).
144. Nowak, M. A. Five rules for the evolution of cooperation. *Science* **314**, 1560 (2006).
145. Axelrod, R. & Hamilton, W. D. The evolution of cooperation.
146. Fehr, E. & Schmidt, K. M. A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics* **114**, 817 (1999).
147. Bó, P. D. Cooperation under the shadow of the future: experimental evidence from infinitely repeated games. *American economic review* **95**, 1591 (2005).
148. Fudenberg, D. & Maskin, E. *The folk theorem in repeated games with discounting or with incomplete information in A Long-Run Collaboration On Long-Run Games* 209 (World Scientific, 2009).

149. Okun, A. M. *Equality and Efficiency – The Big Tradeoff* (The Brookings Institution, 1975).
150. Marden, J. R., Arslan, G. & Shamma, J. S. Joint strategy fictitious play with inertia for potential games. *IEEE Transactions on Automatic Control* **54**, 208 (2009).
151. Gini, C. Variabilità e mutabilità. *Reprinted in Memorie di metodologica statistica* (Ed. Pizetti E, Salvemini, T). Rome: Libreria Eredi Virgilio Veschi (1912).
152. Ceriani, L. & Verme, P. The origins of the Gini index: Extracts from Variabilità e Mutabilità (1912) by Corrado Gini. *The Journal of Economic Inequality* **10**, 421 (2012).
153. Champernowne, D. G. A model of income distribution. *The Economic Journal* **63**, 318 (1953).
154. Riolo, R. L., Cohen, M. D. & Axelrod, R. Evolution of cooperation without reciprocity. *Nature* **414**, 441 (2001).
155. Hetzer, M. & Sornette, D. A theory of evolution, fairness, and altruistic punishment. *PloS ONE* **8**, e77041 (2011).
156. Osborne, M. J. & Rubinstein, A. *A Course in Game Theory* (MIT Press, Cambridge, MA, 1994).
157. Friedman, J. W. A non-cooperative equilibrium for supergames. *The Review of Economic Studies* **38**, 1 (1971).
158. Baumol, W. J. *Welfare Economics and the Theory of the State* (Longmans, Green and co, London, UK, 1952).
159. Alexander, R. D. *The Biology of Moral Systems* (Transaction Publishers, 1987).
160. Mailath, G. J. & Samuelson, L. *Repeated Games and Reputations: Long-run Relationships* (Oxford University Press, Oxford, UK, 2006).
161. Nowak, M. A. & Roch, S. Upstream reciprocity and the evolution of gratitude. *Proceedings of the Royal Society of London B: Biological Sciences* **274**, 605 (2007).
162. Ostrom, E. A behavioral approach to the rational choice theory of collective action: Presidential address, American Political Science Association, 1997. *American Political Science Review* **92**, 1 (1998).
163. Panchanathan, K. & Boyd, R. Indirect reciprocity can stabilize cooperation without the second-order free rider problem. *Nature* **432**, 499 (2004).

164. Fehr, E. Human behaviour: Don't lose your reputation. *Nature* **432**, 449 (2004).
165. Yoeli, E., Hoffman, M., Rand, D. G. & Nowak, M. A. Powering up with indirect reciprocity in a large-scale field experiment. *Proceedings of the National Academy of Sciences* **110**, 10424 (2013).
166. Perc, M., Gómez-Gardeñes, J., Szolnoki, A., Floría, L. M. & Moreno, Y. Evolutionary dynamics of group interactions on structured populations: A review. *Journal of The Royal Society Interface* **10**, 20120997 (2013).
167. Alchian, A. A. & Demsetz, H. Production, Information Costs, and Economic Organization. *American Economic Review* **62**, 777 (1972).
168. Semmann, D., Krambeck, H. J. & Milinski, M. Reputation is valuable within and outside one's own social group. *Behavioral Ecology and Sociobiology* **57**, 611 (2005).
169. Berger, U. & Grüne, A. On the stability of cooperation under indirect reciprocity with first-order information. *Games and Economic Behavior* **98** (2016).
170. Brandt, H. & Sigmund, K. Indirect reciprocity, image scoring, and moral hazard. *Proceedings of the National Academy of Sciences* **102**, 2666 (2005).
171. Panchanathan, K. & Boyd, R. A tale of two defectors: The importance of standing for evolution of indirect reciprocity. *Journal of Theoretical Biology* **224**, 115 (2003).
172. Ohtsuki, H. & Iwasa, Y. How should we define goodness? Reputation dynamics in indirect reciprocity. *Journal of Theoretical Biology* **231**, 107 (2004).
173. Brandt, H. & Sigmund, K. The logic of reprobation: Assessment and action rules for indirect reciprocation. *Journal of Theoretical Biology* **231**, 475 (2004).
174. Cuesta, J. A., Gracia-Lázaro, C., Ferrer, A., Moreno, Y. & Sánchez, A. Reputation drives cooperative behaviour and network formation in human groups. *Scientific Reports* **5**, 7843 (2015).
175. Milinski, M., Semmann, D. & Krambeck, H.-J. Reputation helps solve the 'tragedy of the commons'. *Nature* **415**, 424 (2002).
176. Muller, L., Sefton, M., Steinberg, R. & Vesterlund, L. Strategic behavior and learning in repeated voluntary contribution experiments. *Journal of Economic Behavior & Organization* **67**, 782 (2008).

177. Nax, H. H., Burton-Chellew, M. N., West, S. A. & Young, H. P. Learning in a black box. *Journal of Economic Behavior & Organization* **127**, 1 (2016).
178. Brown, L. D., Cai, T. T. & DasGupta, A. Interval estimation for a binomial proportion. *Statistical Science* **16**, 101 (2001).
179. Grujić, J., Eke, B., Cabrales, A., Cuesta, J. A. & Sánchez, A. Three is a crowd in iterated prisoner's dilemmas: Experimental evidence on reciprocal behavior. *Scientific Reports* **2** (2012).
180. Barcelo, H. & Capraro, V. Group size effect on cooperation in one-shot social dilemmas. *Scientific Reports* **5** (2015).
181. Nosenzo, D., Quercia, S. & Sefton, M. Cooperation in small groups: The effect of group size. *Experimental Economics* **18**, 4 (2015).
182. Fischbacher, U. z-Tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics* **10**, 171 (2007).
183. Bock, O., Baetge, I. & Nicklisch, A. hroot: Hamburg registration and organization online tool. *European Economic Review* **71**, 117 (2014).
184. Mann, H. B. & Whitney, D. R. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics* **18**, 50 (1947).
185. Richardson, A. M. Nonparametric statistics: A step-by-step approach. *International Statistical Review* **83**, 163 (2015).
186. Brown, L. D., Cai, T. T. & DasGupta, A. Interval estimation for a binomial proportion. *Statistical Science*, 101 (2001).
187. Fisher, R. A. The design of experiments. (1935).
188. Good, P. *Resampling Methods: A Practical Guide to Data Analysis* (Birkhäuser, Basel, CH, 1999).
189. Collingridge, D. S. A primer on quantitized data analysis and permutation testing. *Journal of Mixed Methods Research* **7**, 81 (2013).
190. Phipson, B. & Smyth, G. K. Permutation P-values should never be zero: Calculating exact P-values when permutations are randomly drawn. *Statistical Applications in Genetics and Molecular Biology* **9** (2010).
191. Keser, C. & Van Winden, F. Conditional cooperation and voluntary contributions to public goods. *scandinavian Journal of Economics* **102**, 23 (2000).

192. Leifeld, P. *texreg*: Conversion of Statistical Model Output in R to LATEX and HTML Tables. *Journal of Statistical Software* **55**, 1 (2013).
193. Fernihough, A. *Simple logit and probit marginal effects in R* Working Papers 201122 (School of Economics, University College Dublin, 2011).

CURRICULUM VITAE

PERSONAL DATA

Name	Stefano Duca
Date of Birth	December 28, 1988
Place of Birth	Naples, Italy
Citizen of	Italy

EDUCATION

October 2011 – March 2014	Jointly: Ludwig Maximilians Universität and Technische Universität München Munich, Germany <i>Final degree:</i> M.Sc. in Theoretical and Mathematical Physics (1.3/1)
September 2007 – October 2010	Università degli studi di Napoli, Federico II, Naples, Italy <i>Final degree:</i> B.Sc. in Physics (110/110 cum laude)
September 2002 – July 2007	Liceo Scientifico Statale "Leon Battista Alberti", Naples, Italy <i>Final degree:</i> Diploma di Maturità Scientifica (100/100)

EMPLOYMENT

October 2014 –	PhD Student <i>ETH Zürich,</i> Zurich, Switzerland
March 2014 – September 2014	Research Assistant <i>Ludwig Maximilians Universität,</i> Munich, Germany

PUBLICATIONS

Articles in peer-reviewed journals:

1. Duca, S., Helbing, D. & Nax, H. H. Assortative matching with inequality in voluntary contribution games. *Computational Economics* **52**, 1029 (2018).
2. Nax, H. H., Murphy, R. O., Duca, S. & Helbing, D. Contribution-based grouping under noise. *Games* **8**, 50 (2017).
3. Duca, S. & Nax, H. H. Groups and scores: the decline of cooperation. *Journal of The Royal Society Interface* **15**, 20180158 (2018).

Article under review:

4. Duca, S. *Heterogenous agents in voluntary contribution games with assortative matching and wealth accumulation* ETH Zürich Working Paper. 2018.