Challenging Cooperation: Inequality, Global Commons, Future Generations

CHALLENGING COOPERATION:

INEQUALITY, GLOBAL COMMONS, FUTURE GENERATIONS

A dissertation presented

by

Oliver Paul Hauser

to

The Department of Organismic and Evolutionary Biology

in partial fulfilment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Biology

Harvard University

Cambridge, Massachusetts

March 2016

Dissertation Advisor
Martin Andreas Nowak

Author
Oliver Paul Hauser

Challenging Cooperation:

Inequality, Global Commons, Future Generations

Abstract

Cooperation is abundant in the world around us, spanning all levels of biological and social organisation. Yet the existence and maintenance of cooperation is puzzling from an evolutionary perspective because the costs borne to cooperating individuals put them at an evolutionary disadvantage. We thus require an understanding of mechanisms and institutions that can enable cooperation to thrive and be maintained. In this dissertation, I discuss three issues that have presented, or currently present, a challenge to the sustenance of human cooperation. The first chapter addresses an issue of much contemporary debate – inequality. I ask how the well-documented, widespread lack of knowledge of income inequality in society affects the use of costly punishment and costly reward in maintaining public cooperation. When income inequality in a group is not known, the poorest group members are punished (for their low absolute contributions) while the richest are rewarded (for their high absolute contributions). Conversely, when income inequality is revealed, this outcome reverses: the poorest are rewarded (for their high percentage of income contributed) and the richest are punished (for their low percentage contributed). In my next dissertation chapter, I turn to study the emergence of large-scale cooperation. How can cooperation arise and remain stable in large groups? Although it has been argued that the standard reciprocity mechanism weakens in large groups, a simple, scalable intervention—dubbed "local-to-global" reciprocity—successfully maintains public cooperation in groups orders of magnitude larger than previously studied. Local-to-global reciprocity works to maintain group-level cooperation because individuals withhold cooperation from defectors in pairwise interactions as a form of punishment. In the last chapter, I investigate how we can cooperate with future generations: people today face the challenge that they must pay the cost of cooperation now to benefit people in the future who cannot reciprocate their actions. When people decide individually, the renewable resource quickly depletes leaving future generations empty-handed. When decisions today are made by majority vote, however, the resource is sustained for many generations. Voting works because it allows a cooperative majority to restrain a minority of present-day defectors.

Table of Contents

Citations of Previously Published Work

Chapters 1 and 2 are unpublished.

Chapter 3 is published as:

> Hauser O.P.*, Rand D.G.*, Peysakhovich A., Nowak M.A. (2015). Cooperating with the future. *Nature* **511**, 220-223. [*Joint first authors]

Appendix A is published as:

> Hauser O.P., Traulsen A., Nowak M.A. (2014). Heterogeneity in background fitness acts as a suppressor of selection. *Journal of Theoretical Biology* **343**, 178–185.

Acknowledgements

I am grateful to so many people who have helped me along the way to earning this doctorate. Here is my attempt at thanking you, though my gratitude for your help and support can hardly be expressed in a just a few words in print.

I would like to begin by thanking my dissertation advisor Martin Nowak. When I first met Martin, I was an unsuspecting audience member in one of his regular talks on the evolution of cooperation. But the fascination and enthusiasm Martin expressed in this talk, and in the many meetings that were to follow it, irreversibly changed my perspective and, in fact, the course of my doctorate. Martin has been an inspiration from the start and he has taught me what it means to be an academic: his love for his subject, his intuition for asking fascinating questions, and his passion for gaining and understanding knowledge across disciplines are but a few examples of what I have learnt from him over the years. I am grateful to Martin for advising and supporting my decisions, and I look forward to many more years of collaboration and friendship.

I would like to thank the members of my dissertation committee David Haig, Naomi Pierce, and Dave Rand for their advice and support. Dave, in particular, inspired me to work on human cooperation: he taught me how to design experiments and analyse them, how to write them up into cool papers, and he invited me to his lab during my visits to Yale University. I'm also thankful to the faculty and staff at the Organismic and Evolutionary Biology department who made it possible for me to explore new uncharted territory and who supported me throughout this rich, exciting, and often unexpected journey.

Similarly, I am grateful to everyone who has welcomed me into their midst despite my relatively recent transition into the field. My gratitude extends in particular to my friends and mentors at the Harvard Business School. Above all, Mike Norton brought me into HBS and introduced me to life at a business school, kindly welcoming me into his lab on the other side of the river. His enthusiasm, humour, and joy in doing research are a real inspiration in my own work. Mike taught me to be creative in coming up with ideas, to create a friendly environment in which to share them, and the importance of asking "Wouldn't it be cool if …?" in a good research project. I'm also indebted to Max Bazerman and Mike Luca for teaching me how to teach, for advising and mentoring me towards becoming a better experimentalist, and for taking me along with them on their adventures in the HBS/HKS Immersive Field Course. And I'm grateful to the faculty in the Negotiation, Organizations & Markets Unit, including Francesca Gino and Kathleen McGinn, and in the Behavioural Insights Group, including Todd Rogers and Abigail Dalton, for letting me be part of the exciting developments in behavioural science at Harvard.

Next I would like to thank the many wonderful people with whom I have had the pleasure of collaborating and from whom I have learnt so much over the years: Kate Barasz, Bethany Burum, Kate Donahue, Grant Donnelly, Achim Hendriks, Christian Hilbe, Alison Hill, Moshe Hoffman, Anson Kahng, Stewart Kettle, Gordon Kraft-Todd, Madeleine Marino, Johanna Mollerstrom, Alex Peysakhovich, Johannes Reiter, Michael Sanders, Jennifer Sheehy-Skeffington, Arne Traulsen, Carl Veller, and Erez Yoeli.

I'd also like to give a shout-out to my friends and colleagues at the various institutes where I have had the privilege of working over the years. Thank you for welcoming me into your groups and for making my graduate experience so enjoyable in all these different places. At the *Program for Evolutionary Dynamics* at Harvard: Ben Adlam, Ben Allen, Philipp Altrock, Tibor Antal, Peter Blake, Ivana Bozic, Krishnendu Chatterjee, Feng Fu, Katherine Gallagher, Jeff Gerold, Moshe Hoffman, Alison Hill, May Huang, Kamran Kaveh, Scott Kominers, Johanna Mollerstrom, Chuck

Noble, Jason Olejarz, Alex Peysakhovich, Dave Rand, Johannes Reiter, Daria Roithmayr, Daniel Rosenbloom, Sam Sinai, Arne Traulsen, Matthijs van Veelen, Carl Veller, and Erez Yoeli. At the *GiNorton Lab* at the Harvard Business School: Ryan Buell, Grant Donnelly, Erin Frey, Ximena Garcia-Rada, Francesca Gino, Paul Green, Leslie John, Tami Kim, Mike Norton, Patricia Satterstrom, Ovul Sezer, Alison Wood Brooks, and Ting Zhang. At the *Human Cooperation Lab* at Yale: Antonio Alonso Arechar, Adam Bear, Jeremy Cone, Jill Jordan, Matt Jordan, Gordon Kraft-Todd, Alex Peysakhovich, Dave Rand, Jonathan Schulz, Nick Stagnaro, and Christina Starmans. At the *Sprouts Lab* at the Yale School of Management: Daylian Cain, Zoe Chance, Jason Dana, George Newman, Nathan Novemsky, Frederick Shane, and Gal Zauberman. And at the *Behavioural Insights Team* at the UK Cabinet Office: Hannah Behrendt, Raj Chande, Kate Glazebrook, Michael Hallsworth, David Halpern, Sam Hanes, Hugo Harper, Susannah Hume, Miranda Jackman, Andy Jackson, Stewart Kettle, Katy King, Elspeth Kirkman, Ariella Kristal, Elizabeth Linos, Oliver Nguyen, Joanne Reinhard, Simon Ruda, Michael Sanders, Owain Service, Antonio Silva, Zhi Soon, Francesca Tamma, and Alex Tupper.

During my time at Harvard, I have also always been grateful for the many interesting and thought-provoking conversations that inspired new ideas and made graduate life fun, with friends and colleagues from across the university: Netta Barak-Corren, Donal Cahill, Alek Chakroff, Ekin (Dogus) Cubuk, Ben Franta, Jana Gallus, Christof Gebhardt, Benjamin Good, Sioux Hall & Harvard Aikikai, Philipp Harms, Liz Keenan & Aaron Hartmann, Julia Lee, J. Nathan Matias, Eric Mazur, Philip Munoz, Philipp Preiss, Cami Ratliff, Orad Reshef, Mark Anthony Schembri, and Alex Woolf.

Of course, I wouldn't be at Harvard if it had not been for many supportive friends and family in my life. Before coming to Harvard, both my life and my research would certainly not have been half as much fun (but probably twice as productive) without Lismarie Egel, Vanessa Fillafer, Bruno Pfeiffer, and Philipp Waibel. Also in Austria, thank you to my classmates and veteran Innsbruckers Christof Brandtner, Georg Ganglmayr, Max Hanke, Christoph Rauth, Manuel Schischkoff, Patrick Singewald, Lukas Stampfer, and Valentin Torggler; and to Simon Überall and Adham Hamed who have been there from the start. I am deeply grateful to my wonderful family who have been cheering me on from the side lines (read: abroad) with unfailing support. Thank you to Katy, George, Jennifer and Andrew; to my sister Eva; and to my parents Marlis and Dietmar who have always set an example that hard work, joy, and "the good life" can work in harmony.

And, finally, to my wife Emily: I am always grateful for everything you do and I am so fortunate and happy to have you in my life. The fact that we met right at the start of our graduate studies, more than half a decade ago, only goes to show that this wonderful journey has been so much more than "just" earning a doctorate. My life would not be the same without you; nor would my PhD be on the topic that I am truly passionate about. You have been supportive of all that I do, you have helped me come up with new ideas that turned into fascinating research projects, and you have encouraged me to grow both professionally and personally. For your constant belief in me and for your unwavering love – this dissertation is for you!

For Emily, always

Chapter 1.

Introduction

From microbes to humans, cooperation is key to evolutionary success across all levels of biological organisation (Nowak 2006a; Nowak 2006b). Cooperation means incurring a personal cost to provide a benefit to someone else. This poses a challenge for evolutionary biologists to explain the emergence of cooperation in the face of natural selection, the selfish machinery of evolution: natural selection increases the frequency of successful strategies while unsuccessful ones die out. Cooperation, in its pursuit to help others at an individual cost, is thus not favoured by natural selection.

And yet cooperation clearly abounds in the natural world: it ranges from bacteria and fish to ants and birds in the non-primate world, it is supremely prevalent in monkeys and apes, and it is the foundation upon which the lives and societies of humans are built. To explain the evolution of cooperation, five rules (or *mechanisms*) have been proposed and tested experimentally (Nowak 2006b; Dal Bo 2005; Wedekind and Milinski 2000; Rand et al. 2014): direct reciprocity, indirect reciprocity, spatial selection, group selection, and kin selection. Although these mechanisms have been instrumental for our understanding of cooperative behaviour in humans and others species, there remains plenty to be explained. This dissertation, focused on human cooperation, aims to fill a gap in our understanding of "challenging" circumstances for cooperation.

Here, I discuss three major challenges for cooperation. These challenges are merely three vignettes pertaining to the human experience across its evolutionary history to the present day. The challenges discussed here are the effects of inequality in resources upon cooperation, the emergence of large-scale cooperation, and the sustainability of renewable resources for future generations to come. While each of these vignettes has posed a challenge to human cooperation time and again, I believe they can be classified, if only suggestively, as particularly crucial issues that we face today, have faced in the past, or will face in the future. In keeping with this analogy, the first chapter investigates a particularly relevant contemporary issue—inequality—and its consequences on costly punishment and costly reward; the next chapter looks at the emergence of large-scale cooperation, an issue particularly pertaining to early human societies; and the last chapter asks what it takes to cooperate with the future.

In addressing these challenges, I combine a theoretical approach using classical and evolutionary game theory with behavioural human experiments. On the one hand, I show that we can use behavioural experiments to test the theoretical predictions generated by previously proposed models as well as our own models. Classical (or rational) game theory does not always predict human behaviour well (Kahneman 2003; Ariely 2008; Camerer and Fehr 2006); in fact, much of our evidence suggests that an evolutionary, or behavioural, lens can be a more appropriate description of human behaviour in our experiments. The experiments described in Chapter 3 are examples of work deriving from, and testing, evolutionary theory; Appendix A introduces a new theoretical evolutionary framework to stimulate future experimental research. On the other hand, these behavioural experiments in turn feed back into the theoretical literature, prompting the creation of new models and predictions for future research. Chapters 2 and 4 present such experiments where new models will be generated from the empirical evidence.

Chapter 2, entitled *Punishing the poor and rewarding the rich*, is a behavioural experiment studying the effects of visibility of inequality on cooperation, punishment, and reward. Inequality, and the accumulation of wealth across generations, is hardly a new phenomenon, and dates back to early agricultural societies (Mulder et al. 2009). However, many societies have experienced a particularly strong increase in inequality since the industrialisation era and again since the 1970s, and nowhere (in the Global North) was this increase more pronounced than in the United States (Atkinson and Piketty 2007). Recent research suggests, however, that most people in the U.S. and elsewhere are largely unaware of the extent of inequality in their country (Norton and Ariely 2011). What would happen if this veil of ignorance were lifted? And what are the consequences right now of this "invisible" inequality? We conducted repeated Public Goods Games (PGGs) with the possibility of costly punishment, reward, or both after each round. Each participant in a group of five was at the start of the experiment assigned to a place in an income distribution that was derived from the most recent U.S. income distribution by quintiles. We manipulated whether or not participants in our experiment knew the income distribution, and observed participants' contribution and sanctioning behaviour. When income was hidden, we found that participants punished the poor (for their low absolute contributions) and rewarded the rich (for their high absolute contributions). This reversed completely when incomes were revealed: the poor were rewarded (for their high percentage of income contributed) while the rich were punished (for their low percentage contributed). Importantly, revealing incomes had positive effects on sustaining contributions (especially from those with the greatest ability to give) and on reducing end-game inequality among participants. The research presented in Chapter 2 was conducted jointly with Gordon Kraft-Todd, David Rand, Martin Nowak and Michael Norton, and has not yet been published. Kraft-Todd, Norton and I planned the experiment; I performed the experiments, collected the data, and analysed the data; and Kraft-Todd, Rand, Norton, Nowak, and I wrote the paper.

In Chapter 3, entitled *Preserving the global commons*, I present a behavioural experiment that demonstrates how to sustain cooperation at any group size, small or large. At some point in the past, humans made an unprecedented leap from small-scale societies, in which a handful of people lived together in disparate tribes, to large-scale, agriculture-based and, more recently, industrialised societies covering the entire planet (Henrich et al. 2010). What made cooperation at this large scale possible? Some authors have argued that reciprocity-based mechanisms cannot explain the emergence of large-scale cooperation (Boyd and Richerson 1988): cooperation is after all no longer stable when interactions move from dyadic relationships to three or more people in a group (Grujić et al. 2012). What this line of argument fails to recognise, however, is that interactions in large groups can be—and, in fact, *are* in the real world—coupled with frequent dyadic interactions. For example, the neighbour who goes the extra mile to help out in a community project today is still your neighbour tomorrow: would you turn her away if she came asking for help the next day? Repeated interactions between pairs of individual allow for targeting reward at cooperators for good behaviour towards the group (and, conversely, for targeting punishment at defectors). Therefore, we designed a behavioural experiment, in which we manipulated the key aspect to sustaining cooperation in a large group – the ability to observe the contributions to the public good made by one’s neighbours. We coupled a repeated two-stage economic game: in the group contribution stage, all participants in our experiment could contribute in a large-scale PGG that benefitted everyone; in the pairwise cooperation stage, participants played a repeated Prisoner’s Dilemma (PD) with the same individuals (their “neighbours”) over the course of the experiment. In one of the largest public goods experiments to date (with over 1,000 people playing the same PGG), we found that contributions in the treatment condition, where participants could see their neighbours’ PGG contribution, were sustained over time, while contributions decayed over time in the control condition, where participants did not know how much their neighbours contributed in the PGG. We further characterised the mechanism sustaining cooperation in the treatment condition: participants "punished" low-contributing neighbours by withholding cooperation from them in the PD, which subsequently led those neighbours to increase their PGG contributions in the future. The work presented in Chapter 3 was conducted jointly with Achim Hendriks, David Rand, and Martin Nowak, and has not yet been published. Rand, Nowak and I planned the study; Hendriks and I performed the experiments and collected the data; I performed the data analysis; and Rand, Nowak and I wrote the paper.

Chapter 4, entitled *Cooperating with the future*, presents a behavioural experiment comparing mechanisms to successfully sustain a pool of resources between groups over time. Cooperating with the future requires making sacrifices today. However, unlike in other public goods games (Rand et al. 2009; Fehr and Gächter 2002; Ostrom 1990), future generations cannot reciprocate our actions today. What mechanisms can maintain cooperation with the future? To answer this question, we devised a new economic game, called the "Intergenerational Goods Game" (IGG): a line-up of successive groups (generations) can each extract resources from the pool or leave something for the next group. If the group exhausts the pool beyond a (commonly known) threshold, the pool depletes completely and leaves all future generations empty-handed; otherwise it refills and the next group faces the same dilemma. When decisions were made individually, we found that most pools were destroyed quickly. This failure to cooperate with the future was driven by a minority of defectors who extracted far more than what was sustainable. To address this inefficiency, we introduced a social institution that has become a hallmark of many human societies today: democracy. When decisions were decided by median voting (Holcombe 1989), where each individual in the acting generation proposed an amount to extract and the median proposal was implemented for everyone, the resource was consistently sustained across many generations. What makes voting successful is (*i*) the ability of a majority of cooperators to restrain defectors and (*ii*) the reassurance for conditional cooperators (Fischbacher, Gächter, and Fehr 2001) that their efforts are not futile. The mechanisms that make voting work in the first place thus generated a new hypothesis: voting is only successful if it is binding for everyone. Indeed, we found that groups in which three out of five members voted on their decision while the other two made their decisions individually, were not able to sustain cooperation with the future. In fact, bootstrapping simulations based on our results demonstrate that "partial voting" does no better than no voting at all. Thus, for cooperation with the future to work effectively, decisions need to be binding and made as a collective. The work presented in Chapter 4 was conducted jointly with David Rand, Alexander Peysakhovich, and Martin Nowak, and was published in *Nature* in May 2014. Rand, Peysakhovich, Nowak and I planned the study; I performed the experiments and collected the data; Rand and I analysed the data; and Rand, Peysakhovich, Nowak and I wrote the paper.

Finally, I present an evolutionary game theoretic model of inequality in Appendix A, entitled *Heterogeneity in background fitness acts as a suppressor of selection*. This chapter is not focused on cooperation but instead looks at the spread of evolutionary strategies more generally. In particular, we seek to understand the effect of inequality, or heterogeneity, on fixation. While heterogeneity has been introduced to spatial structure, networks, and types of interactions, fitness advantages that are inherent to an individual (not an evolutionary strategy), or to a specific location used for breeding and reproduction, has not been considered. Combining an analytic Markov chain approach with agent-based evolutionary simulations, we introduced heterogeneity into the fitness function $f_i = b_i + s_i$ where f is the net fitness of individual i which is made up of individual background fitness *b* and the payoff derived from playing strategy *s*. We found that inequality in background fitness acts as a suppressor of new strategies in the population. In particular, we identified that, when a certain amount of wealth or resources is added to a population, the more unequal the distribution of allocated resources, the more selection is suppressed. We analytically calculated a strategy's fixation probability under background heterogeneity in small populations. In addition, we found a simple analytical approximation that holds for small and medium-sized populations. The work in Appendix A was conducted jointly with Arne Traulsen and Martin Nowak, and was published in the *Journal of Theoretical Biology* in January 2014. Traulsen, Nowak, and I planned the project; Traulsen and I performed the analytical calculations; I conducted the agent-based simulations; and Traulsen, Nowak and I wrote the paper.

Chapter 2.

Punishing the poor and rewarding the rich

2.1 Main text

Preferences for spending on public goods such as social programmes or the health care system are based, at least in part, on beliefs about income and wealth inequality (Alesina and Angeletos 2005; Charité, Fisman, and Kuziemko 2015; Durante, Putterman, and van der Weele 2014; Kuziemko et al. 2015). Yet recent cross-cultural evidence suggests that across the globe, many people are unaware of the true extent of inequality in their country (Norton and Ariely 2011; Kiatpongsan and Norton 2014; Davidai and Gilovich 2015).

Here, we explore the implications of this lack of awareness of inequality, examining how hiding inequality affects societal outcomes and people's behaviours towards the rich and the poor (relative to when inequality is revealed). We hypothesised that this lack of inequality information could have negative impacts on societal well-being: if people do not realise how little the poor have, and how much the rich have, they may be less sympathetic to low contributions from those who cannot afford to give more, and less likely to hold the rich to account for not contributing their "fair share."

To explore these predictions experimentally, we used a standard paradigm in experimental economics; an incentive-compatible, repeated public goods game (PGG) in groups of 5 players. In each of 10 rounds, every player was assigned an "income" and chose how much of that income to contribute to a common pool; all contributions were doubled and divided equally among the five players (see Section 2.2.2 for more details about study design). We then showed each player the contributions of all other players (player IDs were shuffled every round), and gave each player the opportunity for costly sanctioning of all other players. In the *punishment* condition, participants could pay 1 unit to decrease any other participant's payoff by 3 units; in the *reward* condition, participants could pay 1 unit to increase any other participant's payoff by 3 units. Participants could spend up to 2 units on each other participant.

Such sanctioning schemes have been widely used in previous work on cooperation (Fehr and Gächter 2002; Nikiforakis and Normann 2007; Rand et al. 2009; Sefton, Shupp, and Walker 2007; Sutter, Haigner, and Kocher 2010; Gächter, Renner, and Sefton 2008). In many Western societies, results typically reveal that low contributors are punished, while high contributors are rewarded; anti-social punishment aimed at high contributors, on the other hand, is frequently observed in countries with a weak rule of law (Herrmann, Thöni, and Gächter 2008). In these studies, all players typically receive identical endowments in each round, and this equality is common knowledge to all players; thus the majority of these experiments, while highly informative regarding the maintenance of cooperation, shed little light on perceptions of, and reactions to, inequality.

Recently, however, scholars have been investigating inequality in the provisioning of public goods. In these experiments, richer participants have been found to contribute less of their income than poorer ones, decreasing overall social efficiency (Buckley and Croson 2006; Isaac and Walker 1988; Keser et al. 2014; Keser, Markstädter, and Schmidt 2014). When punishment is introduced, participants punished others to express their preference for efficiency norms (everyone contributing their entire income) or relative contribution norms (everyone contributing the same relative share of their income), even when such behaviour led to reduction in average net earnings

(Reuben and Riedl 2013; Antinyan, Corazzini, and Neururer 2015; Gächter et al. 2014). Moreover, uncertainty around others' actual contributions or incomes also elicited greater punishment (Ambrus and Ben Greiner 2012; Bornstein and Weisel 2010). (See Section 2.2.1.2 for details on previous literature.)

Building on this previous research, we introduce three novel features to explore the impact of people's recently demonstrated lack of knowledge regarding the distribution of income (Norton and Ariely 2011; Kiatpongsan and Norton 2014): (*i*) we experimentally vary whether the income distribution, as well as each participant's specific income, is hidden or revealed to explore the causal effect of knowledge of inequality on behaviour toward the rich and poor; (*ii*) we use an income distribution that is extremely unequal (the actual United States distribution) to explore behaviour toward the rich and poor under conditions reflective of real-world inequality; and (*iii*) we allow participants to either punish, reward, or both punish and reward the poor and the rich to explore how these sanctions are utilised to address perceived inequity. Indeed, we expected sanctions to be crucial in addressing inequality and net payoffs. Absent sanctions, visibility of wealth inequality can actually increase inequality in social networks (Nishi et al. 2015). But in the presence of sanctions, we predicted that revealing incomes would be a solution, not an obstacle: when incomes are revealed, the ability to reward and punish allows participants to shape others' future contributions and restore equity.

Across both experiments, we used the United States pre-tax incomes by quintile to create incomes for the five players. In our first experiment, the top quintile participant received 55 units out of 100 units in the group (or 55% of all income), the next 19 units, the next 13 units, the next 9 units, and the bottom quintile participant 4 units (Fig. 1a) (Congressional Budget Office 2007). Once assigned to an income level, participants received the same income each round for 10 rounds.

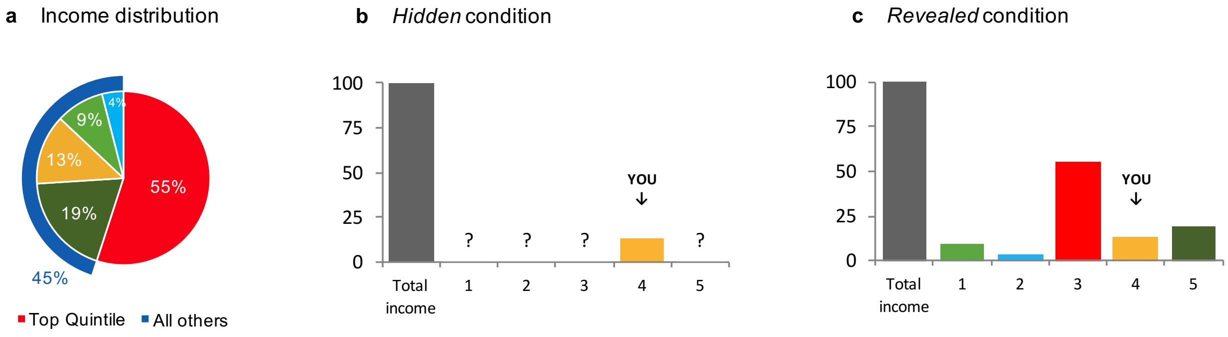

***Figure 2.1. The income distribution in our game and the main experimental manipulation between the* hidden *and* revealed *conditions. a** Each player in a group of five was randomly assigned to a position in an income distribution. In the first experiment, we used the 2007 U.S. pre-tax income distribution* (Congressional Budget Office 2007)*: in each of ten rounds, the top quintile participant received 55 units, while the bottom quintile player received 4 units. **b** When making decisions to punish and reward, participants in the* hidden *condition saw their own income and the sum of all incomes. **c** In the* revealed *condition, participants viewed all players' incomes.*

The design of our first experiment was a 2 (*punishment* versus *reward*) X 2 (*hidden* versus *revealed*) between-participants design (*N* = 600). In the *hidden* condition, players had no information about the incomes of the others in their group (Fig. 1b): they made contributions, viewed others' contributions, and decided to punish or reward based only on the total amounts contributed by other players. In the *revealed* condition, in contrast, participants were shown the income of each player as they made their decisions to punish or reward – allowing them to base their decisions not only on the total amount contributed, but also the *percentage* of available income that each player chose to contribute (Fig. 1c). For example, a player who contributed just three units in the *hidden* condition may look stingy; learning that this player had only four total units in the *revealed* condition may dramatically alter perceptions of their contribution.

We expected that in the *hidden* condition, participants would generally view the (low total) contributions of bottom quintile players unfavourably, inducing punishment, and the (high total) contribution of the top quintile players favourably, inducing reward. In contrast, we expected that in the *revealed* condition, participants would generally view the (high percentage) contributions of bottom quintile players favourably, inducing reward, and the (low percentage) contribution of the top quintile players unfavourably, inducing punishment.

We find that, indeed, participants in the *hidden* condition rewarded richer participants more (coeff = 0.636, $p < 0.001$), whereas those in the *revealed* condition rewarded poorer participants more (coeff = -0.720, $p < 0.001$; interaction between income and *revealed* dummy, coeff = -1.356, $p < 0.001$; Figure 2.2 and Table 2.1). We observe a mirror image of these results for decisions to punish: participants in the *hidden* condition punished poorer participants more (coeff = -0.282, $p = 0.042$), whereas those in the *revealed* condition punished richer subjects more (coeff = 0.692, $p < 0.001$; interaction between income and *revealed* dummy, coeff = 0.974, $p < 0.001$; Figure 2.2 and Table 2.3). Thus, knowledge about economic inequality had a profound effect on sanctioning. (Unless otherwise noted, all statistical analyses use linear regression with standard errors clustered on group, taking income quintile as a continuous predictor variable; note that using log-transformed income instead of income quintile as a predictor generates qualitatively equivalent results, see Section 2.2.2.)

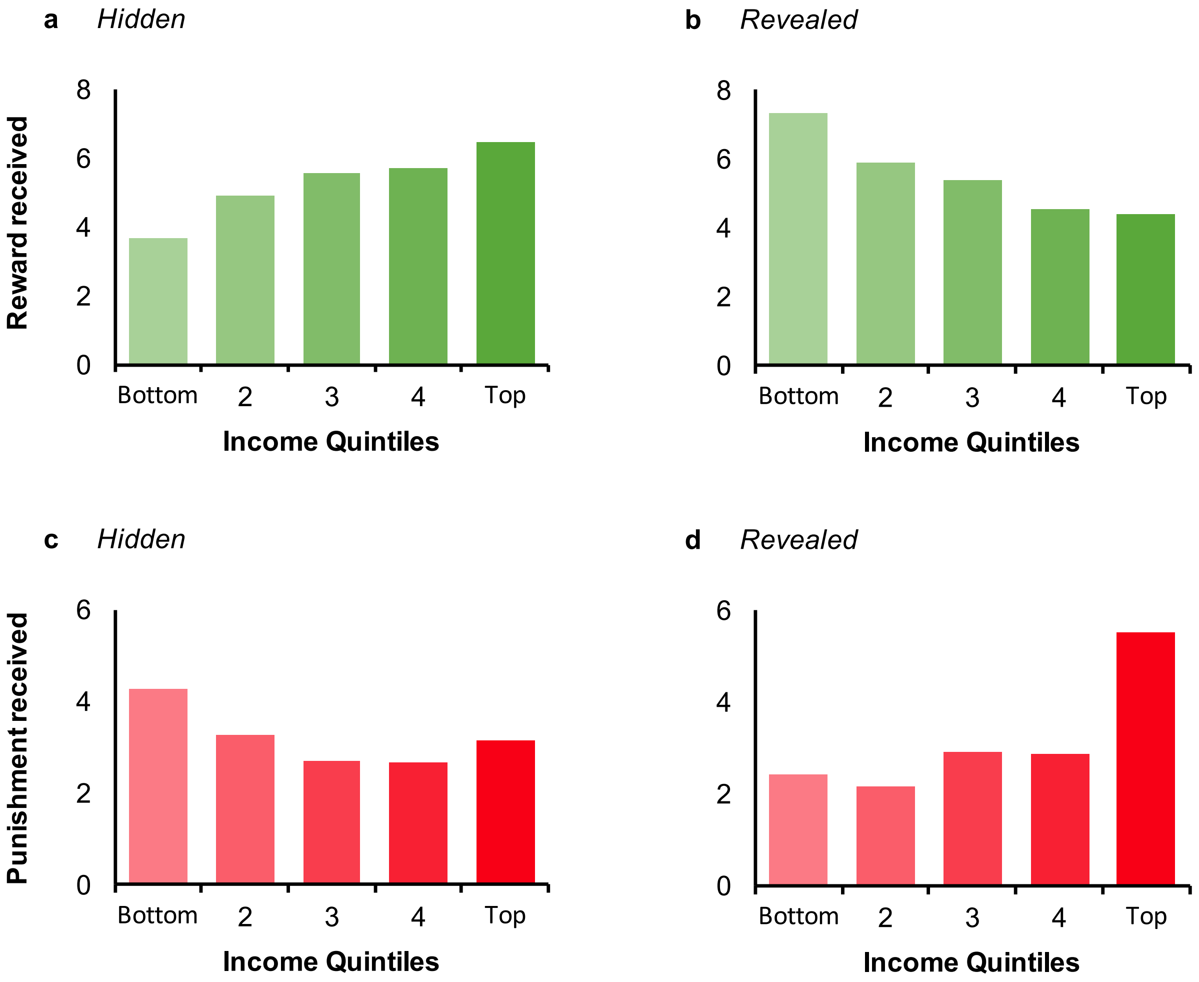

Figure 2.2. Amount of received reward (top panels) and punishment (bottom panels) depends on income quintile and whether income was hidden (left panels) or revealed (right panels). a *Participants rewarded higher income participants more in the* hidden *condition, but* ***b*** *less in the* revealed *condition.* ***c*** *Punishment behaviour is a mirror image of reward: participants punished poorer participants more in the* hidden *condition, while* ***d*** *punishing richer participants more in the* revealed *condition.*

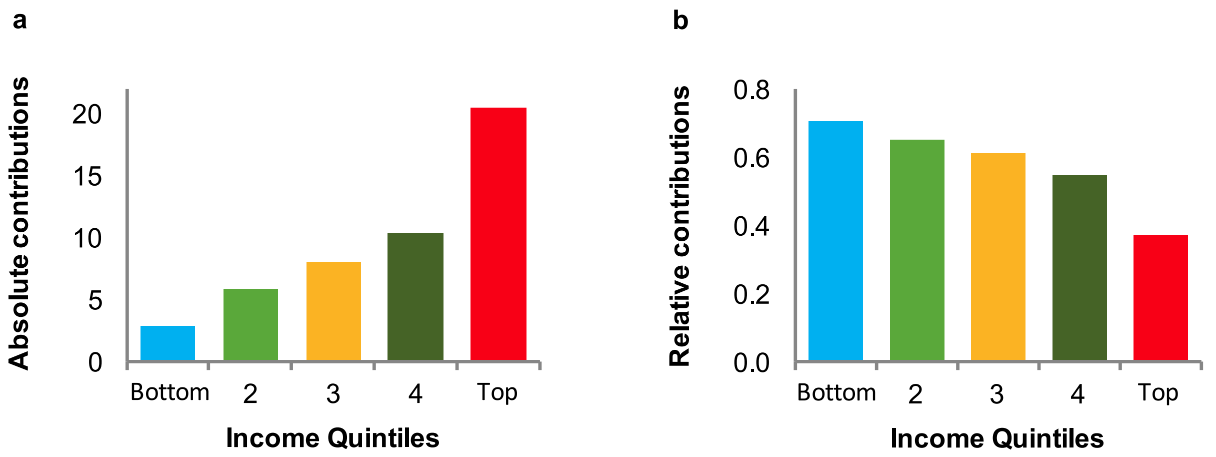

Figure 2.3. Who contributes more? a *In the* hidden *condition, only absolute contributions could be assessed, such that richer participants appeared to contribute more.* ***b*** *In the* revealed *condition, where participants could view contributions relative to income, it was clear that lower income participants contributed a larger fraction of their income.*

Why did players sanction so differently in the *hidden* and *revealed* conditions? Across both conditions, richer players contributed larger total amounts (*hidden*: coeff = 3.172, $p < 0.001$; *revealed*: coeff = 4.734, $p < 0.001$; Table 2.5), but lower percentages of their income (*hidden*: coeff = -0.098, $p < 0.001$; *revealed*: coeff = -0.058, $p < 0.001$; Table 2.6) (Figure 2.3). Collapsing across conditions, top quintile participants contributed 20.49 out of 55 units (or 37% of their income) whereas bottom quintile participants contributed 2.83 out of 4 units (or 71% of their income). The pattern of sanctioning we observe therefore follows naturally if sanctions were assigned based on percentage of income contributed in the *revealed* condition but total amount contributed in the *hidden* condition.

Supporting this logic, in the *revealed* condition, participants conditioned their sanctioning decisions on the percentage of the target's income that was contributed (using percentage contributed as the independent variable; predicting punishment, coeff = -4.664, p < 0.001, Figure 2.4a; predicting reward, coeff = 6.320, p < 0.001, Figure 2.4b; Table 2.10), more so than on the absolute amount contributed (if anything, absolute contribution predicts the opposite direction from the overall observed pattern: using log-transformed total contribution as the independent variable; predicting punishment, higher absolute contributors are punished more, coeff = 1.365, p = 0.003; predicting reward, higher absolute contributors are rewarded marginally less, coeff = -0.980, p = 0.080; Table 2.10). In fact, in the *revealed* condition, participants not only rewarded and punished based on percentage contributed; they also targeted the richest players in particular, even when they contributed the same relative amount of their income (using percentage contributed and income as the independent variables; predicting punishment, coeff on quintile = 0.563, p < 0.001; predicting reward, coeff on quintile = -0.309, p = 0.031; Tables 2.11 and 2.12). Thus, sanctions were used partly to encourage future contributions from the rich and partly to reduce the wealth gap in the group.

In the *hidden* condition, conversely, where only total contribution amounts were known, sanctioning was based on total amount contributed (using log-transformed total contribution as the independent variable; predicting punishment, coeff = -1.863, p = 0.019, Figure 2.4c; predicting reward, coeff = 4.700, p < 0.001, Figure 2.4d; Table 2.9), but not on percentage of income contributed (using percentage contributed as the independent variable; predicting punishment, coeff = 0.030, p = 0.954; predicting reward, coeff = -0.216, p = 0.677; Table 2.9).

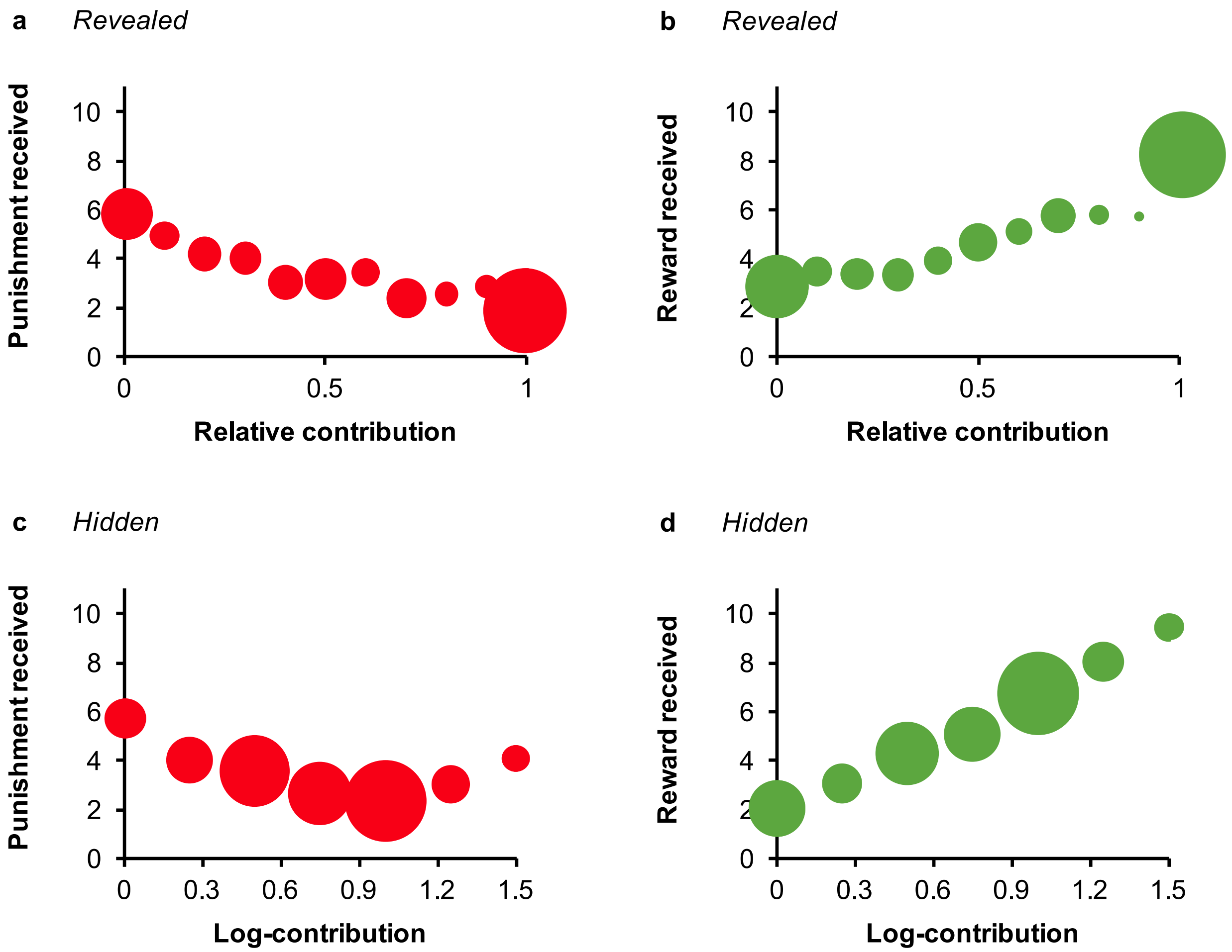

***Figure 2.4. Received punishment and reward depends on percentage of income contributed in the revealed condition (top panels) and on absolute income contributed in the hidden condition (bottom panels). a,b** When incomes were revealed, participants who contributed a higher percentage of their income were **a** punished less ($p < 0.001$) and **b** rewarded more ($p < 0.001$). **c,d** Conversely, when incomes were hidden, participants who contributed a higher absolute amount were **c** punished less ($p = 0.042$) and **d** rewarded more ($p < 0.001$). Bubble size is proportional to the fraction of corresponding participants.*

We next consider the consequences of income transparency on total contributions. Overall, significantly more units were contributed in the *revealed* condition compared to the *hidden* condition (coeff = 1.745, $p = 0.002$; Table 2.15). Participants in the bottom (poorest) through fourth (second richest) quintiles maintained (or even increased) their contribution levels over the ten rounds in both the *hidden* and *revealed* conditions (no significant change in contribution over round, $p = 1.00$ bonferroni-corrected for all quintile-condition pairs, with the exception of the poorest quintile in the *hidden* condition, who actually increased over time: coeff = 0.047, $p = 0.04$ corrected; Tables 2.17 and 2.18). However, although participants in the top (richest) quintile in the *revealed* condition also continued to contribute over time (coeff = -0.382, $p = 1.000$ corrected), top quintile players in the *hidden* condition decreased their contributions over the ten rounds (coeff = -1.077, $p < 0.001$ corrected) (Tables 2.17 and 2.18). Thus, in the *hidden* condition, sanctions were less effective at maintaining contributions among those with the greatest ability to contribute to the public good.

Participants in our initial experiment were assigned their income randomly. However, incomes in the real world are not just the product of chance, but also of effort. In a second experiment, we assigned income based on participants' performance in an individual effort task before playing the public goods game. The best-performing participant in a group earned the highest income, the 2nd-best performing participant earned the 2nd-highest income, and so on.

In addition, we further increased the external validity of our results. We told participants (who had been recruited exclusively from the U.S.) that the income distribution used in the game was derived from the U.S. income distribution. Thus, if participants had an accurate estimate of U.S. inequality, they would be less likely to 'punish the poor' in the *hidden* condition. In addition, we gave participants a wider range of simultaneous actions towards other participants in the second experiment: they could choose to reward or punish different participants in each round, paying 2 unit to increase or decrease another participant's payoff by 6 units, respectively. (For details about experiment 2, see Section 2.2.2.2.)

We found qualitatively similar results in our second experiment: participants continued to reward the rich and punish the poor in the *hidden* condition (predicting number of units received: coeff = 0.053, *p* = 0.042), while this trend reversed completely in the *revealed* condition (coeff = -0.171, *p* < 0.001; interaction between income and *revealed* dummy: coeff = -0.225, *p* < 0.001; Table 2.20). Across conditions, richer participants contributed more in absolute terms (coeff = 3.979, p < 0.001), but less as a percentage of their income (coeff = -0.078, p < 0.001), than poorer participants, a pattern linked to reward and punishment decisions: higher absolute contributions received more reward in the *hidden* condition (coeff = 0.571, p < 0.001, Table 2.26), but higher percentage of income contributed was more rewarded in the *revealed* condition (coeff = 1.775, p < 0.001, Table 2.26). Even when incomes were earned *and* when participants were informed that the income distribution was reflective of their own country's distribution, participants continued to punish the poor and reward the rich when the income distribution was hidden, but reward the poor and punish the rich when incomes were revealed.

In sum, revealing inequality had substantial effects on people's decisions to reward or punish others, and on total contributions to the public good. Participants were more likely to punish poorer participants and reward richer participants when inequality was hidden; when income was revealed, participants became more sensitive to people's *ability* to contribute – leading them to punish the rich and reward the poor.

While income and wealth heterogeneity has long fascinated theorists and experimentalists alike, only recently have the effects of high levels of inequality representative of many real countries been considered (Olson 1965; Baland and Platteau 1997; Dieckmann and Kun 2013; Hauser, Traulsen, and Nowak 2014). Here, we have shown experimentally that even high levels of inequality need not hinder contributions to the public good (Olson 1965), but *lack* of awareness of inequality can impede the ability of sanctions to sustain the commons – especially in soliciting contributions from the rich.

While our income distribution was drawn from the real world, our paradigm necessarily offers a stylised examination of the impact of inequality on the public good. For example, we restricted the amount that all participants could pay to punish or reward other players to 2 units per player. The real world may not always provide such an upper bound: given their greater resources, the rich have much greater ability to inflict harm or bestow benefits on others. Still, there are some real-world situations in which all decisions count equally: for instance, casting a vote in democratic elections carries equal weight despite differences in income.

Revealing incomes decreased inequality and increased total contributions in our experimental groups, with implications for policymakers concerned with the public good. While revealing all citizens' incomes may seem challenging to implement—or even hard to imagine—in some countries, it is common practice in others. For example, the Norwegian government operates an online database which contains detailed information about all citizens' income, wealth, and tax contributions (Norwegian Tax Administration 2015). Notably, Norwegians also have very high tax morale (I. Lago-Peñas and Lago-Peñas 2010); while anecdotal, this example suggests that revealing incomes can be associated with increased support for the public good – mirroring our results. In a world of income transparency, the "haves" may become more generous and the "have-nots" less punished, with positive implications for the common good.

2.2 Supporting figures and data

2.2.1. Motivation and relation to previous work

2.2.1.1 Research motivation

The central observation motivating the present work is that people, both in the U.S. and in 39 other countries, systematically underestimate the extent of inequality in their country (Norton and Ariely 2011; Kiatpongsan and Norton 2014). We hypothesised that this misinformation regarding inequality can have major negative impacts on societal well-being: if people do not realise how little the poor have, and how much the rich have, they may be less sympathetic to low contributions from those who cannot afford to give more, and less able to hold the rich to account for not contributing their “fair share”.

To explore these societal impacts experimentally, we used the standard paradigm from experimental economics for studying group social interactions: the public goods game (PGG). In particular, we built on the large body of prior work demonstrating that people tend to punish players who do not contribute in the PGG, and reward players who do (Fehr and Gächter 2000; Fehr and Gächter 2002; Rand et al. 2009; Sefton, Shupp, and Walker 2007; Sutter, Haigner, and Kocher 2010; Herrmann, Thöni, and Gächter 2008; Gächter, Renner, and Sefton 2008). (There are, however, cultural differences across countries: while punishment of low contributors is the norm in most Western countries, so-called ‘anti-social’ punishment aimed at high contributors is observed at high frequency in countries with a weak rule of law (Herrmann, Thöni, and Gächter 2008).) While most prior PGG studies have focused on groups where incomes were equally distributed, several recent studies have begun to explore how inequality affects contribution and sanctioning behaviour (Buckley and Croson 2006; Reuben and Riedl 2013) (see Section 2.2.1.2 for more details).

In the current paper, we add to our understanding of inequality by incorporating three key features of inequality that have received little prior attention. First, in most prior work, the income distribution was common knowledge among all PGG groups members. Thus, little is known about our central question of the consequences of the empirical observation that people do not have an accurate understanding of the level of inequality (Norton and Ariely 2011; Kiatpongsan and Norton 2014). To that end, we experimentally manipulate when the income distribution is known or hidden.

Secondly, most prior studies focus on the impact of levels of inequality that were much lower than what is observed outside the laboratory. The Gini index is the most common measure of inequality (Allison 1978). Using the most recent country-level data from the World Bank (World Bank, n.d.), we found that globally, the Gini index ranges between 0.25 and 0.66. Almost no prior studies used endowments that reflected that level of inequality: while 90% of all countries had Gini indices higher than 0.29, 91% of previous PGG experiments we surveyed had a Gini index below 0.29. To better reflect the reality of income inequality, we used PGG endowments that match the actual U.S. income quintiles from 2007 having a Gini index of 0.440 (Experiment 1), and 2013 having a Gini index of 0.444 (Experiment 2).

Third, most prior studies have randomly assigned subjects to receive higher or lower incomes. Yet in reality, variation in income is (at least in part) determined by non-random factors such as effort. Thus we also explore the impact of earned vs random inequality, and how this interacts with knowledge about the distribution of incomes.

Finally, while most prior experiments have either not allowed for sanctioning, or have focused exclusively on punishment, we examine both costly peer punishment and costly peer rewarding. Rewards (e.g. "positive" sanctions) play a key role in much of human social life, and thus understanding whether inequality impacts rewards differently from punishment is important.

2.2.1.2 Previous research

The vast majority of prior literature using public goods game has focused on equally-endowed participants. In recent years, however, economists and psychologists have begun to study the effects of endowment (or income) inequality on cooperation. Here we provide a brief survey of prior literature on inequality and public goods. (While we focus on *income* inequality, we note that inequality can also be induced by varying the marginal per capita return (Fisher et al. 1995; Reuben and Riedl 2008; Cardenas, Stranlund, and Willis 2002), show-up fees for participants (Anderson, Mellor, and Milyo 2008), or marginal abatement costs (Brick and Visser 2012), or by taking advantage of endogenous variation in participants' real-world wealth (Cardenas 2003; Cardenas 2007).)

One of the most consistent findings has been that exogenous inequality in incomes leads to lower levels of overall contributions in a group (Isaac and Walker 1988; Buckley and Croson 2006; Keser et al. 2014; Cherry, Kroll, and Shogren 2005). The reduction in contributions is driven primarily by richer participants contributing less so as to match the level of contributions of the poorer participants (Buckley and Croson 2006). Consequently, wealthier participants contribute less relative to their income than do poorer participants (Buckley and Croson 2006; Chan et al. 1996; Keser et al. 2014).

Several researchers have aimed to introduce interventions to increase cooperation between unequal group members. Institutional fines for low contributors (Brick and Visser 2012), minimum contribution requirements (Keser, Markstädter, and Schmidt 2014), communication (Chan et al. 1999; Hackett, Schlager, and Walker 1994; Brick and Visser 2012) and punishment (Bornstein and Weisel 2010; Reuben and Riedl 2013; Antinyan, Corazzini, and Neururer 2015) can play an important role in sustaining contributions.

Of particular relevance for the current paper, Reuben and Riedl (2013) demonstrated that punishment can stabilise contributions from participants with different incomes. In their experiments, participants used punishment in order to enforce both efficiency norms (everyone contributing their entire income) as well as relative contribution norms (everyone contributing the same fraction of their income); Carpenter and Matthews (2009) found similar results among equally-endowed players (Carpenter and Matthews 2009). Similarly, Antinyan et al. (2013) and Bornstein and Weisel (2010) showed that punishment was effective in sustaining contributions in unequal groups under fully informed condition. (Note that while Bornstein and Weisel (2010) also investigate a situation in which players' incomes in each round were not observable, the *distribution* of incomes is always known in their experiment, and each subject typically has experience receiving each possible level of income.)

Importantly, research using PGGs without inequality shows that incomplete information about contributions can lead to increased spending on punishment (Ambrus and Ben Greiner 2012) and lower average payoffs (Grechenig and Nicklisch 2010). Thus, it seems likely that prior conclusions about punishment and inequality may change when the income distribution is unknown.

Furthermore, some research indicates that the origin of incomes can matter. Most studies randomly assign incomes to participants. However, various theories of fairness suggest that earning incomes by exerting individual effort can lead to more acceptance of inequality. The experimental literature has led to mixed results on this conjecture: Van Dijk and Wilke (1994) found that participants who were made to believe that their group members earned more money by exerting more effort were more likely to contribute more to the public good, and vice versa (Van Dijk and Wilke 1994).

In contrast, Cherry et al. (2005), Hofmeyr et al. (2007) and Antinyan et al. (2015) found that public goods contributions were not significantly different when endowments were earned or received as a 'windfall' (Cherry, Kroll, and Shogren 2005; Hofmeyr, Burns, and Visser 2007; Antinyan, Corazzini, and Neururer 2015). In some cases, punishment towards low contributors can be reduced when income was earned (Antinyan, Corazzini, and Neururer 2015). Thus, understanding how earned incomes affects cooperation and sanctioning, and how it might interact with (lack of) knowledge of the income distribution, remains an open question.

2.2.2. Methods

2.2.2.1 Data collection on Amazon Mechanical Turk

We recruited U.S. residents to participate using the online labour market Amazon Mechanical Turk (AMT). AMT is an online market place in which employers can pay users for completing short tasks (generally about 10 minutes) – usually referred to as Human Intelligence Tasks (HITs) – for a relatively small payment (generally less than a $1). Workers who have been recruited on AMT receive a baseline payment and can be paid a bonus depending on their performance in the task. This setup lends itself to incentivised economic experiments: the baseline payment acts as the 'show up' fee and the bonus payment may derive from the workers' behaviour in the economic game and/or other tasks throughout the experiment.

The sample of recruited participants on AMT has been shown to be more diverse and more nationally representative than subject pools at most research universities (Buhrmester, Kwang, and Gosling 2011). Numerous studies have been carried out to validate results collected using AMT

(Berinsky, Huber, and Lenz 2012; Crump, McDonnell, and Gureckis 2013; Paolacci and Chandler 2014). Of particular relevance are studies showing quantitative agreement between play in economic games conducted on AMT and in the physical laboratory (Mason and Suri 2011; Horton, Rand, and Zeckhauser 2011; Amir et al. 2012; Rand, Greene, and Nowak 2012).

All data was collected using Software Platform for Human Interaction Experiments (SoPHIE) (Hendriks 2012). Experiment 1 was carried out in summer 2014, while Experiment 2 was conducted in summer 2015. SoPHIE is a novel experimental platform that enables participants to interact with one another in real time. Participants were recruited on the AMT website, were grouped together, and then made decisions simultaneously; their decisions were exchanged through an external server provided by SoPHIE Labs (www.sophielabs.net).

The experiments were approved by the Harvard University Committee on the Use of Human Subjects in Research.

2.2.2.2 Basic flow of the experiments

2.2.2.2.1 Experiment 1: ‘windfall’ incomes

All participants earned a $1.50 showup fee and had the opportunity to earn additional bonus payments between $0.00 and $3.88, depending on the outcome of the game. Participants took part in the experiment through an online survey provided by SoPHIE Labs. After participants had read the experiment instructions (see Section 2.2.4), they had to pass a comprehension quiz about the rules of the game in order to partake in the actual experiment. Participants who did not pass the comprehension quiz on the first attempt were given the chance to try again; no participants were thus removed from the experiment. After participants passed all comprehension questions, they waited for up to 10 minutes in a designated online ‘waiting room.’ As soon as five participants had arrived in the waiting room, the public goods game (PGG) started automatically. A step-by-step account from the participants' perspective in each condition can be found in Section 2.2.4.

We recruited at least 30 groups of five participants who completed the experiment in each condition ($N = 600$; including drop-out groups, $N = 855$). At the beginning of the game, participants were randomly assigned to a position in the income distribution. This implied that participants earned a 'windfall' endowment, which can in some cases affect contribution behaviour in public goods games (Van Dijk and Wilke 1994). In experiment 2, we show, however, that assigning incomes based on performance, instead of randomly assigning incomes, does not alter our reward and punishment results.

The income distribution was common knowledge to all participants in the *revealed* condition, but not in the *hidden* condition. Across conditions, the actual distribution (Congressional Budget Office 2007) is the same: the participant with the highest income (referred to as top quintile player) earns 55 units per round, the second-highest earner 19 units, the middle participant 13 units, the second-lowest 9 units and the bottom quintile player 4 units. That is, the sum of income units distributed each round is 100 units, and this total is also known to all players in all conditions (see Section 2.2.4 for screenshots). Once incomes had been assigned, participants received the same income in each of the 10 rounds.

The game lasted 10 rounds and each round consisted of two stages. Participants were given no information about the length of the game to avoid any end-round effects, as in prior work (Rand et al. 2009; Rand et al. 2014). In stage 1, participants were asked how many units they wanted to contribute to the public good. All contributed units were doubled and then split equally among all five players. In stage 2, participants were either participating in a punishment-game or a reward-game. In the punishment game, each player could pay 1 unit to take away 3 units from another player. In the reward game, conversely, players could pay 1 units to increase another player's payoff by 3 units.

Across conditions, we limited the number of units that could be spent on punishment and reward to 2 per target player. In addition, no participant could spend more than they had accumulated with their income and their earnings from stage 1. In other words, participants could spend up to 8 units per round on the other 4 players but if they had less than 8 units in their account, the upper bound of spending was their remaining endowment. To examine whether this upper bound could affect the decisions participants were able to make (e.g., consistently prevented them from punishing or rewarding), we examined all players' payoffs after stage 1 across all conditions. We find that only 0.84% of the time did a participant have less than 8 units available, and the number of affected participants does not vary by condition ($\chi^2(21) = 23.16$, $p = 0.34$). Given the small number of incidences in which participants were constrained in their decisions to punish and reward, it is unlikely that they affect our results.

At the end of each round, participants were informed about the payoff they earned. At the end of 10 rounds, participants filled out a short demographics survey. If at any time a participant became unresponsive (because they quit the game or lost their Internet connection), the remaining participants in the group were automatically moved to an 'early exit' screen. They were informed that a participant had left the game unexpectedly and thus the experiment could not be continued. All remaining participants were asked to fill out a questionnaire and earn a bonus of $1.00 to compensate them for their time spent on the study.

Across all conditions, in 29.8% of groups, a participant dropped out before the end of the game (which is consistent with previous studies carried out on AMT, e.g. (Rand, Greene, and Nowak 2012)). Dropout rates did not vary between conditions ($\chi^2(3) = 2.74$, $p = 0.433$). In half of the groups that dropped out, one or more participants did not respond within the time available to them (up to two minutes per decision stage). In the other half, one of the participants in each dropout group quit their browser or tab (either by choice or due to a failed Internet connection). Dropout rates did not differ significantly between quintiles ($\chi^2(12) = 7.89$, $p = 0.793$). The majority of dropouts (73.9%) occurred in the first half of the game and the time of dropout did not differ by quintile (χ^2 (21) = 23.123, $p = 0.337$).

We further analysed participants' likelihood of dropping out based on the average sum of contributions in the group or the rewards and punishment a participant received. Across conditions, the sum of contributions did not affect whether a participant finished the game (logistic regression using sum of a group's contributions to predict finishing the game: coeff = 0.007, $p = 0.340$). There was not interaction between the sum of contributions and income visibility (logistic regression using sum of contributions interacted with *revealed* dummy: coeff = 0.008, $p = 0.614$).

The amount of reward or punishment received did not affect a participant's likelihood of finishing the game (logistic regression using sanctions received to predict finishing the game; reward: coeff = 0.023, $p = 0.618$; punishment: coeff = -0.048, $p = 0.093$). Income visibility did not affect dropout rates: the amount a participant was rewarded or punished did not significantly affect their chances of finishing the game (logistic regression using sanctions received interacted with *revealed* dummy; reward: coeff = 0.135, $p = 0.101$; punishment: coeff = -0.028, $p = 0.600$).

In our main analysis, we include groups that did not finish the game. However, we found qualitatively similar results when we dropout groups are not included in the analysis.

2.2.2.2.2 Experiment 2: earned incomes

All participants earned a $2.00 show-up fee and had the opportunity to earn an additional in bonus payments between $0.00 and $4.26 depending on the outcome of the game. Participants took part in the experiment through SoPHIE. The experiment consists of two phases.

In phase 1, participants completed an individual task (Abeler et al. 2011). They were asked to count the number of 0s in a matrix randomly made up of 0s and 1s. The goal was to complete as many such matrices as possible within 2 minutes. Each time, after they submitted a solution, a new matrix was presented to them. Participants were not informed of how many matrices they had solved correctly. However, they were told that their performance mattered for their income in the upcoming group task. The best performing participant in the individual task was assigned the highest income level; the second-best performing participant received the second-highest level of income, and so on. Participants were fully informed in both conditions about the assignment procedure of incomes. (See Section 2.2.4 for screenshots of the instructions.)

In phase 2, participants then read instructions about an economic game. After reading the instructions, the were asked to answer several comprehension questions. Once participants had correctly answered the comprehension questions, they waited for four other participants to arrive in the 'waiting room'. As soon as five participants were ready, the began the repeated two-stage economic game comprising of a public goods stage and a sanctioning stage.

We collected 30 groups of five participants who completed the game in each condition (*N* = 300; including dropout groups, *N* = 440). We excluded participants from experiment 1 from participating in experiment 2. While conceptually similar, the economic game in experiment 2 differed from the economic game in experiment 1 in several ways. First, the income distribution used in experiment 2 was updated to reflect the latest U.S. census data (U.S. Census Bureau 2013):

the top quintile player received 51 units, the player in the next quintile 23 units, the middle-quintile player 15 units, the second-lowest player 8 units and the lowest earner 3 units (Figure 2.5).

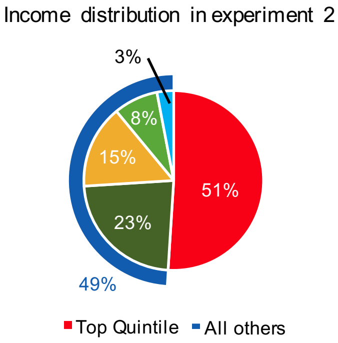

Figure 2.5. *The income distribution used in experiment 2: the income quintiles were derived from the latest available U.S. income distribution* (U.S. Census Bureau 2013).

In addition, participants in experiment 2 knew that the income distribution used in this game represented the current income distribution of the United States. Unlike in experiment 1, participants thus had an implicit reference point, which they could use to make informed reward and punishment decisions assuming that their estimate of U.S. income inequality is accurate. Participants in experiment 1 were not given any hint about the distribution used and only knew that their incomes would be different from one another. Based on previous research (Norton and Ariely 2011), our prediction was that participants in experiment 2 would, however, misestimate the extent of U.S. inequality and it would thus qualitatively not alter the way in which people punished or rewarded.

Furthermore, the reward and punishment conditions in experiment 2 were collapsed into one combined treatment: it was thus possible for participants to punish or reward in all conditions in experiment 2. Each player could choose not to pay any units and thus leave another player's payoff unaffected; or they could choose to pay 2 units to reduce another player's by 6 units or pay 2 units to increase another player's by 6 units. Participants could spend up to 8 units on all four other players or up to as much as they had earned in stage 1. Only 2.33% of the time a participant was restricted in punishing or rewarding their group members because they had less than 8 units available in stage 2, which did not vary by condition ($\chi^2(9) = 12.532$, $p = 0.185$).

Finally, we adopted the standard infinite-game paradigm used in economics (Dal Bo 2005): participants were told that the game would last at least 8 rounds and each additional round would occur with a probability of 50% to avoid end-game effects (Dal Bo 2005; Rand et al. 2009). Due to a programming error, we did not collect demographic information from participants in experiments 2. However, since participants were randomly assigned to the *hidden* or *revealed* condition, there should not be any systematic variation in demographics across conditions. Furthermore, when we controlled for demographics in experiment 1, we found that it did not affect our results.

Dropout rates in experiment 2 were comparable to those in experiment 1: 31.82% of groups did not finish the game in experiment 2. Dropout rates did not differ by condition ($\chi^2(1) = 0.098$, $p = 0.755$). Across quintiles, there was no variation in the rate of dropout ($\chi^2(4) = 3.280$, $p = 0.512$). The majority of dropouts (71.4%) occurred in the first five rounds of the game; and the time of dropout did not differ by condition ($\chi^2(4) = 4.278$, $p = 0.370$).

The behavioural history of participants did, however, matter to whether a group finished the game in experiment 2. While the sum of contributions across both condition did not affect a participant's likelihood of finishing (logistic regression using sum of group contributions to predict finishing the game: coeff = 0.013, $p = 0.112$), there was an interaction between the sum of contributions and income visibility: participants in the *revealed* condition were more likely to finish the game, the more the group contributed to the common pool, while there was no effect of sum of contributions in the *hidden* condition (logistic regression using sum of contributions as IV to predict finishing the game: coeff = -0.014, p = 0.186; sum of contributions interacted with *revealed* dummy: coeff = 0.043, p = 0.004).

The amount of reward or punishment participants received also affected their dropout rates in experiment 2. Across both conditions, participants were not significantly affected by rewards and punishment received (logistic regression using number of units received to predict finishing the game: coeff = 0.180, p = 0.094). However, a difference emerged by condition: whereas groups in the *revealed* condition were more likely to finish the game when they received more rewards, whereas there was no effect of units received in the *hidden* condition (logistic regression using number of units received to predict finishing the game: coeff = -0.168, p = 0.287; sum of contributions interacted with *revealed* dummy: coeff = 0.66, p < 0.001).

Since neither sum of contributions nor received reward or punishment predicted dropout rates in the *hidden* condition, we can be reasonably confident that selection effects are not driving our results in the *hidden* condition. However, in the *revealed* condition, groups in which fewer contributions were made altogether and groups in which fewer reward units were distributed were more likely to drop out. Thus, to mitigate concerns about potential selection effects, we include all groups in our main analyses, regardless of whether they completed the game or dropped out. However, we found qualitatively similar results when we did not include dropout groups.

2.2.3. Statistical details

2.2.3.1 Experiment 1

Unless otherwise noted, all statistics are linear regressions with income quintile as a continuous independent variable. Because participants' decisions are not independent within each group over time, we cluster standard errors on the group. To check the robustness of our results, we also use log-transformed absolute income as continuous IV. We use log-transformed income, rather than absolute income, because the distribution of incomes is highly right skewed (the income of the rich participant is an outlier relative to the other 4 income levels).

2.2.3.1.1 Received reward and punishment

We first assessed whether income visibility affects participants' reward and punishment behaviour. We examined which player(s) participants rewarded and punished when not informed about the income distribution (*hidden* condition) compared to when they were informed (*revealed* condition). The independent variable in our main analysis was the income quintile of the recipient of the sanctions (1 to 5). The higher the participant's quintile, the higher her income: the participant in the first quintile was the poorest player (1 = bottom quintile) whereas the participant in the fifth quintile was the richest player (5 = top quintile).

We found qualitatively similar results when we used amount of income of the recipient as the independent variable, rather than quintile. To account for the fact that the distribution of incomes is highly right skewed (the income of the top quintile player is an outlier), we used log-transformed income amounts. We included the regression table of the log-transformed income models below each of the corresponding quintile models.

2.2.3.1.1.1 Reward

In the *hidden* condition, participants could not take another player's ability to contribute to the public good into account, since this information is not available to them. We thus expected participants to view the (high total) contributions of the top quintile participants favourably—leading to more reward targeted toward them. In the *revealed* condition, conversely, we predicted that participants would view the (high percentage) contributions of the bottom quintile participants favourably, inducing higher reward.

To test these hypotheses, we estimated the amount of reward that a participant receives as a function of their income quintile and whether the income distribution was hidden or revealed (Table 2.1). In the *hidden* condition, we found that higher income participants were rewarded significantly more (coeff = 0.636, $p < 0.001$, Table 2.1 col. 1), whereas in the *revealed* condition, lower income players were rewarded significantly more (coeff = -0.720, $p < 0.001$, Table 2.1 col. 2). Furthermore, a regression including data from both *hidden* and *revealed* conditions together showed that this difference was itself significant (interaction between income and *revealed* dummy, coeff = -1.356, $p < 0.001$; Table 2.1).

We also found qualitatively similar results when we included demographic information: higher income participants were rewarded more in the *hidden* condition (coeff = 0.629, $p < 0.001$, Table 2.1 col. 4) while, conversely, they were rewarded less in the *revealed* condition (interaction between income and *revealed* dummy: coeff = -1.320, $p < 0.001$, Table 2.1 col. 4).

Finally, we repeated the same analysis with log-transformed income as independent variable. We found qualitatively similar results: higher income participants were rewarded more in *hidden* (coeff = 1.043, $p < 0.001$, Table 2.2 col. 1) but, conversely, they were rewarded less in *revealed* (coeff = -1.140, $p < 0.001$, Table 2.2 col. 2), and the interaction between condition and income was significant (interaction between log-transformed income and *revealed* dummy, coeff = -2.183, $p < 0.001$, Table 2.2 col. 3). The results were qualitatively similar when demographics were included (Table 2.2 col. 4).

Table 2.1: Linear regression model estimating the effect of a target's income quintile (i.e., their position in the income distribution) on the amount of reward they received. Standard errors clustered on group.

	(1)	(2)	(3)	(4)
VARIABLES	*Hidden*	*Revealed*	Interaction	Interaction
Quintile	0.636***	-0.720***	0.636***	0.629***
	(0.141)	(0.134)	(0.141)	(0.143)
1=Revealed			4.316***	4.180***
			(0.957)	(0.961)
Quintile X Revealed			-1.356***	-1.320***
			(0.193)	(0.199)
1=Female				0.561
				(0.468)
Age				0.0368
				(0.0279)
Location				-0.809
				(0.490)
Constant	3.332***	7.648***	3.332***	2.119*
	(0.534)	(0.801)	(0.531)	(0.917)
Observations	1,735	1,690	3,425	3,381
R-squared	0.038	0.038	0.039	0.051

Robust standard errors in parentheses

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

Table 2.2: Linear regression model estimating the effect of a target's log-transformed income on the amount of reward they received. Standard errors clustered on group.

	(1)	(2)	(3)	(4)
VARIABLES	*Hidden*	*Revealed*	Interaction	Interaction
Log(income)	1.043***	-1.140***	1.043***	1.014***
	(0.223)	(0.226)	(0.221)	(0.228)
1=Revealed			5.966***	5.749***
			(1.130)	(1.142)
Log(income) X Revealed			-2.183***	-2.111***
			(0.315)	(0.328)
1=Female				0.553
				(0.470)
Age				0.0370
				(0.0276)
Location				-0.799
				(0.489)
Constant	2.509***	8.475***	2.509***	1.345
	(0.632)	(0.945)	(0.628)	(0.976)
Observations	1,735	1,690	3,425	3,381
R-squared	0.038	0.036	0.037	0.049

Robust standard errors in parentheses

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

2.2.3.1.1.2 Punishment

We followed the same analysis procedure as above for punishment. We expected a mirror image of the results compared to reward. In the *hidden* condition, we predicted that participants would view the (low total) contributions of the bottom quintile participants unfavourably, leading them to punish them more. In the *revealed* condition, participants were able to take the participant's ability to contribute into account and thus we expected that the (low percentage) contributions of the top quintile participants would be viewed disapprovingly, leading to higher punishment.

In our analysis, the dependent variable was the amount of punishment that a participant received. The independent variables are the income quintile and whether the income distribution was hidden or revealed (Table 2.3). In the *hidden* condition, we found that higher income participants were punished less (coeff = -0.282, *p* = 0.042, Table 2.3 col. 1), whereas in the *revealed* condition, in contrast, higher income participants were punished more (coeff = 0.692, *p* < 0.001, Table 2.3 col. 2). Furthermore, a regression including data from both *hidden* and *revealed* conditions together showed that this difference was itself significant (interaction between income and *revealed* dummy, coeff = 0.974, *p* < 0.001; Table 2.3 col 3).

Qualitatively similar results were obtained when we included demographics in the regression: higher income participant in the *hidden* condition were punished marginally less (coeff = -0.258, *p* = 0.057), while, in the *revealed* condition, higher income participants were punished more (interaction between quintile and *revealed* dummy: coeff = 1.005, *p* < 0.001).

We found qualitatively similar results with log-transformed income: higher income participants were punished marginally less in *hidden* (coeff = -0.417, *p* = 0.057, Table 2.4 col. 1) but were punished more heavily in *revealed* (coeff = 1.212, *p* < 0.001, Table 2.4 col. 2); a difference which was itself significant (interaction between log-transformed income and *revealed* dummy, coeff = 1.629, p < 0.001, Table 2.4 col. 3). Results are qualitatively similar when demographics were included (Table 2.4 col. 4).

Table 2.3: Linear regression model estimating the effect of a target's income quintile on the amount of punishment they received. Standard errors clustered on group.

	(1)	(2)	(3)	(4)
VARIABLES	*Hidden*	*Revealed*	Interaction	Interaction
Quintile	-0.282*	0.692***	-0.282*	-0.258
	(0.133)	(0.0982)	(0.132)	(0.134)
1=Revealed			-2.956***	-3.045***
			(0.640)	(0.665)
Quintile X Revealed			0.974***	1.005***
			(0.164)	(0.170)
1=Female				-0.185
				(0.273)
Age				0.0129
				(0.0153)
Location				0.185
				(0.307)
Constant	4.056***	1.099**	4.056***	3.614***
	(0.513)	(0.391)	(0.509)	(0.687)
Observations	1,590	1,819	3,409	3,230
R-squared	0.011	0.061	0.038	0.045

Robust standard errors in parentheses

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

Table 2.4: Linear regression model estimating the effect of a target's log-transformed income on the amount of punishment they received. Standard errors clustered on group.

	(1)	(2)	(3)	(4)
VARIABLES	*Hidden*	*Revealed*	Interaction	Interaction
Log(income)	-0.417	1.212***	-0.417	-0.376
	(0.212)	(0.168)	(0.210)	(0.212)
1=Revealed			-4.303***	-4.459***
			(0.798)	(0.828)
Log(income) X Revealed			1.629***	1.690***
			(0.269)	(0.277)
1=Female				-0.182
				(0.276)
Age				0.014
				(0.015)
Location				0.181
				(0.306)
Constant	4.302***	-0.001	4.302***	3.788***
	(0.635)	(0.493)	(0.630)	(0.796)
Observations	1,590	1,819	3,409	3,230
R-squared	0.009	0.070	0.042	0.050

Robust standard errors in parentheses

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

2.2.3.1.2 Absolute vs. relative contribution

What caused participants to punish and reward other players so differently in the *hidden* and *revealed* conditions? Contribution behaviour provides a potential answer. We hypothesised that the *hidden* and *revealed* conditions enabled participants to view contributions differently: in the *hidden* condition, participants could not take another player's ability to contribute into account since they did not know the income distribution. In the *revealed* condition, on the other hand, participants could evaluate the amount contributed relative to the player's income before choosing whom to punish or reward – in other words, participants could differentiate between absolute and relative contributions.

The hypothesis that relative contributions were driving the difference in sanctions between *hidden* and *revealed* generated several predictions. First, we expected that absolute contributions would be higher for higher income participants while relative contributions (as a percentage of income) would be higher for lower income participants.

Second, we expected absolute contributions to predict received reward and punishment when income is hidden, but relative contributions to predict received reward and punishment when income is revealed: participants would punish (reward) those who give little (a lot) in absolute terms more in *hidden*, while their sanctions would be driven by relative contributions in *revealed*.

2.2.3.1.2.1 Absolute vs. relative contribution by quintile

We first examined absolute contributions by quintile. We expected that higher income participants contributed a larger amount of units to the public good but a smaller fraction of their total income – such that lower income participants would contribute a larger percentage of their income.

As predicted, we found that higher income participants in both the *hidden* (coeff = 3.172, $p < 0.001$, Table 2.5 col. 1) and *revealed* (coeff = 4.738, $p < 0.001$, Table 2.5 col. 3) conditions contributed a larger number of units. Conversely, we found that higher income participants made smaller relative contributions (percentage of income contributed) in both *hidden* (coeff = -0.098, $p < 0.001$, Table 2.6 col. 1) and *revealed* (coeff = -0.058, $p < 0.001$, Table 2.6 col. 3). All results are robust to inclusion of demographic variables (Tables 2.5 and 2.6 cols. 2 and 4).

This variation in contribution across incomes can also be illustrated with an example: across all conditions, top quintile participants contributed 20.49 out of 55 units (or 37% of their income) to the public good. In contrast, bottom quintile participants contributed 2.83 out of 4 units (or 71% of their income). In Tables 2.7 and 2.8, we repeated the same analysis with log-transformed income as the independent variable; results were qualitatively similar.

Table 2.5: Linear regression model estimating the effect of income on absolute contribution. Standard errors clustered on group.

	(1)	(2)	(3)	(4)
VARIABLES	*Hidden*	*Hidden*	*Revealed*	*Revealed*
Quintile	3.172***	3.195***	4.734***	4.644***
	(0.257)	(0.260)	(0.341)	(0.349)
1=Female		-1.285*		0.367
		(0.645)		(0.830)
Age		0.065		-0.025
		(0.038)		(0.035)
Location		0.381		0.304
		(0.571)		(0.752)
Constant	-0.923	-2.550	-3.853***	-3.097*
	(0.476)	(1.338)	(0.637)	(1.314)
Observations	3,412	3,343	3,655	3,461
R-squared	0.233	0.241	0.316	0.312

Robust standard errors in parentheses

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

Table 2.6: Linear regression model estimating the effect of income on *percentage* of income contributed (relative contribution). Standard errors clustered on group.

	(1)	(2)	(3)	(4)
VARIABLES	*Hidden*	*Hidden*	*Revealed*	*Revealed*
Quintile	-0.098***	-0.097***	-0.058***	-0.061***
	(0.009)	(0.009)	(0.010)	(0.011)
1=Female		-0.050		0.025
		(0.027)		(0.029)
Age		0.002		-0.000
		(0.001)		(0.001)
Location		0.053		0.030
		(0.032)		(0.036)
Constant	0.850***	0.788***	0.773***	0.762***
	(0.032)	(0.055)	(0.036)	(0.059)
Observations	3,412	3,343	3,655	3,461
R-squared	0.141	0.153	0.047	0.053

Robust standard errors in parentheses

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

Table 2.7: Linear regression model estimating the effect of log-transformed income on absolute contribution. Standard errors clustered on group.

	(1)	(2)	(3)	(4)
VARIABLES	*Hidden*	*Hidden*	*Revealed*	*Revealed*
Log(income)	5.416***	5.455***	8.082***	7.937***
	(0.464)	(0.467)	(0.611)	(0.626)
1=Female		-1.292*		0.472
		(0.617)		(0.785)
Age		0.060		-0.023
		(0.037)		(0.034)
Location		0.531		0.218
		(0.551)		(0.734)
Constant	-5.598***	-7.155***	-10.827***	-10.068***
	(0.918)	(1.539)	(1.211)	(1.639)
Observations	3,412	3,343	3,655	3,461
R-squared	0.253	0.261	0.344	0.339

Robust standard errors in parentheses

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

Table 2.8: Linear regression model estimating the effect of log-transformed income on *percentage* of income contributed. Standard errors clustered on group.

VARIABLES	(1) *Hidden*	(2) *Hidden*	(3) *Revealed*	(4) *Revealed*
Log(income)	-0.163***	-0.160***	-0.101***	-0.105***
	(0.014)	(0.015)	(0.016)	(0.017)
1=Female		-0.050		0.023
		(0.027)		(0.030)
Age		0.002		-0.000
		(0.001)		(0.001)
Location		0.049		0.031
		(0.032)		(0.036)
Constant	0.981***	0.914***	0.863***	0.857***
	(0.042)	(0.063)	(0.047)	(0.067)
Observations	3,412	3,343	3,655	3,461
R-squared	0.144	0.156	0.052	0.059

Robust standard errors in parentheses

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

2.2.3.1.2.2 Absolute vs. relative contribution predicted sanctioning behaviour

Our hypothesis was that sanction behaviour followed from the contribution pattern. In the *hidden* condition, participants could only consider absolute contribution: thus, because richer participants contributed more units, they would be rewarded more and punished less. In the *revealed* condition, conversely, participants could consider the amount contributed relative to the amount players earn—that is, the percentage of income participants contributed. Because poorer participants contributed more relative to their income in the in the *revealed* condition, they would receive more reward and less punishment.

To test these hypotheses, we examined the effect on sanctioning of the target's relative contribution (percentage of income contributed) and absolute contribution. We log-transformed absolute contribution because of the same right skew that also underlies absolute income, and we added 1 to all contributions prior to log-transforming as the log(0) is undefined (Rand, Greene, and Nowak 2012; McDonald 2014).

In the *hidden* condition, as predicted, we found that higher absolute contributions led to less punishment (coeff = -1.863, $p = 0.019$, Table 2.9 col. 1) and more reward (coeff = 4.700, $p < 0.001$, Table 2.9 col. 3) received. Because richer participants contributed a larger absolute number of units, they received less punishment and more reward in *hidden*. Relative contributions, in contrast, predicted neither punishment ($p = 0.954$, Table 2.9 col. 1) nor reward ($p = 0.677$, Table 2.9 col. 4) received in the *hidden* condition; this is unsurprising, given that relative contributions were not observable in the *hidden* condition. We found qualitatively similar results when including demographics (Table 2.9 cols. 2 and 4).

In the *revealed* condition, participants could assess both relative and absolute contributions, and we expected them to primarily pay attention to relative contribution. Indeed, we found that a higher percentage of income contributed led to less punishment (coeff = -4.664, $p < 0.001$, Table 2.10 col. 1) and more reward (coeff = 6.782, $p < 0.001$, Table 2.10 col. 4) received. This is in line with our prediction: as we have shown before, poor participants contributed a larger percentage of their income and are thus punished less and rewarded more in the *revealed* condition.

We also observed an effect of absolute contribution in the *revealed* condition, in the opposite direction of the effect in the *hidden* condition: higher absolute contributions led to more punishment (coeff = 1.365, $p = 0.003$, Table 2.9 col. 1) and marginally less reward (coeff = -0.980, $p = 0.080$, Table 2.9 col. 4). Although richer participants contributed a larger amount of units, they made low relative contributions; because those larger absolute contributions were correlated with the lowest relative contributions, larger absolute contributions were punished more and rewarded less.

Table 2.9: Linear regression model estimating the effect of absolute log-transformed contribution and relative contribution on received punishment (cols. 1 and 2) and reward (cols. 3 and 4) in the *hidden* condition. To deal with zero-contributions, a constant of 1 was added to all contributions before applying the log-transformation. Standard errors clustered on group.

	(1)	(2)	(3)	(4)
VARIABLES	Punishment received	Punishment received	Reward received	Reward received
Log(contribution+1)	-1.863*	-1.730*	4.700***	4.592***
	(0.755)	(0.739)	(0.659)	(0.641)
Relative contribution	-0.030	-0.256	-0.216	-0.112
	(0.515)	(0.522)	(0.514)	(0.518)
1=Female		-0.243		0.711
		(0.306)		(0.366)
Age		0.021		0.018
		(0.021)		(0.023)
Location		0.272		-0.719
		(0.402)		(0.400)
Constant	4.774***	4.154***	1.624**	0.960
	(0.565)	(0.835)	(0.475)	(0.818)
Observations	1,590	1,553	1,735	1,713
R-squared	0.033	0.037	0.188	0.198

Robust standard errors in parentheses

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

Table 2.10: Linear regression model estimating the effect of relative contribution and absolute log-transformed contribution on received punishment (cols. 1 and 2) and reward (cols. 3 and 4) in the *revealed* condition. To deal with zero-contributions, a constant of 1 was added to all contributions before applying the log-transformation. Standard errors clustered on group.

VARIABLES	(1) Punishment received	(2) Punishment received	(3) Reward received	(4) Reward received
Relative contribution	-4.664***	-4.782***	6.320***	6.420***
	(0.489)	(0.532)	(0.993)	(1.034)
Log(contribution+1)	1.365**	1.465**	-0.980	-0.987
	(0.434)	(0.471)	(0.545)	(0.601)
1=Female		-0.213		0.036
		(0.401)		(0.692)
Age		0.011		0.050
		(0.019)		(0.036)
Location		0.073		-1.542*
		(0.403)		(0.759)
Constant	4.889***	4.636***	2.641***	1.351
	(0.363)	(0.613)	(0.466)	(0.994)
Observations	1,819	1,677	1,690	1,668
R-squared	0.122	0.128	0.182	0.205

Robust standard errors in parentheses

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

2.2.3.1.2.3 Social preference or a desire to take from the rich?

We hypothesised that the poor (rich) would be rewarded (punished) in the *revealed* condition because of their high (low) relative contributions. Next, we were interested in whether reward or punishment behaviour in the *revealed* condition was motivated by more than just eliciting higher relative contributions, such as potentially a desire to target and reduce the income of the rich.

To evaluate this prediction, we tested whether both relative contribution and income quintile both predicted punishment and reward received in the *revealed* condition. Indeed, holding constant the fraction of income contributed, richer participants were rewarded less (coeff = -0.309, *p* = 0.031, Table 2.11 col. 1) and punished more (coeff = 0.563, *p* < 0.001, Table 2.12 col. 1), indicating that participants were not only concerned with the higher income participant's relative contribution but generally more willing to take units from, and less willing to give units to, the rich.

Results were qualitatively similar when demographics are included. Holding constant relative contribution, higher quintiles were rewarded marginally less (coeff = -0.298, *p* = 0.052, Table 2.11 col. 2) and punished more (coeff = 0.604, *p* < 0.001, Table 2.12 col. 2). Similar results are obtained with log-transformed income as IV (Table 2.13 and S14).

While our results shed light on the combined effect of relative contribution and income quintile, it remains an important question for future research to explore to what extent income rank alone motivates the targeting of sanctions.

Table 2.11: Linear regression model estimating the effect of relative contribution and quintile on received reward in the *revealed* condition. Standard errors clustered on group.

	(1)	(2)
VARIABLES	Reward received	Reward received
Relative contribution	5.209***	5.318***
	(1.038)	(1.031)
Quintile	-0.309*	-0.298
	(0.138)	(0.148)
1=Female		-0.016
		(0.695)
Age		0.049
		(0.036)
Location		-1.549*
		(0.757)
Constant	3.426***	2.146
	(0.727)	(1.102)
Observations	1,690	1,668
R-squared	0.184	0.207

Robust standard errors in parentheses

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

Table 2.12: Linear regression model estimating the effect of relative contribution and income quintile on received punishment in *revealed*. Standard errors clustered on group.

	(1)	(2)
VARIABLES	Punishment received	Punishment received
Relative contribution	-3.224***	-3.212***
	(0.377)	(0.394)
Quintile	0.563***	0.604***
	(0.096)	(0.103)
1=Female		-0.181
		(0.402)
Age		0.010
		(0.019)
Location		0.133
		(0.414)
Constant	3.501***	3.107***
	(0.455)	(0.593)
Observations	1,819	1,677
R-squared	0.147	0.156

Robust standard errors in parentheses

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

Table 2.13: Linear regression model estimating the effect of relative contribution and income quintile on received reward in *revealed*. Standard errors clustered on group.

	(1)	(2)
VARIABLES	Reward received	Reward received
Relative contribution	5.239***	5.346***
	(1.042)	(1.038)
Log(income)	-0.441	-0.428
	(0.234)	(0.253)
1=Female		-0.011
		(0.697)
Age		0.050
		(0.036)
Location		-1.547*
		(0.757)
Constant	3.639***	2.350
	(0.907)	(1.193)
Observations	1,690	1,668
R-squared	0.182	0.205

Robust standard errors in parentheses

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

Table 2.14: Linear regression model estimating the effect of relative contribution and income quintile on received punishment in *revealed*. Standard errors clustered on group.

VARIABLES	(1) Punishment received	(2) Punishment received
Relative contribution	-3.162***	-3.136***
	(0.380)	(0.393)
Log(income)	0.983***	1.060***
	(0.162)	(0.172)
1=Female		-0.173
		(0.402)
Age		0.012
		(0.019)
Location		0.117
		(0.408)
Constant	2.577***	2.045**
	(0.553)	(0.620)
Observations	1,819	1,677
R-squared	0.152	0.162

Robust standard errors in parentheses

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

2.2.3.1.3 Public good provisioning and inequality

We next explored the effect of revealed and hidden incomes on public good provisioning as well as subsequent inequality. We find that revealing incomes had a positive effect on total contributions to the public good that was provided – and from whom these contributions came.

2.2.3.1.3.1 Revealing incomes increased contributions

We assessed the effect of revealing incomes on total contributions and whether certain players in the income distribution were affected more than others. Overall, contributions were higher in the *revealed* than in the *hidden* condition (coeff = 1.745, *p* = 0.002, Table 2.15 col. 1). Examining how income and condition interact, we saw that in the *hidden* condition, higher income participants contributed more than lower income participants (coeff = 3.172, *p* < 0.001, Table 2.15 col. 2); and that this difference became significantly larger when incomes were revealed (interaction between income and *revealed* dummy, coeff = 1.562, *p* < 0.001, Table 2.15 col. 2). We found qualitatively equivalent results when demographics (Table 2.15 col. 3) are included as well as when log-transformed income is used as the independent variable (Table 2.16).

Table 2.15: Linear regression model estimating the effect of income visibility (*revealed* dummy) and income on average contribution to the public good. Standard errors clustered on group.

	(1)	(2)	(3)
VARIABLES	Contribution	Contribution	Contribution
1=Revealed	1.745**	-2.930***	-2.651**
	(0.562)	(0.793)	(0.817)
Quintile		3.172***	3.190***
		(0.256)	(0.257)
Quintile X Revealed		1.562***	1.434**
		(0.426)	(0.437)
1=Female			-0.412
			(0.533)
Age			0.016
			(0.026)
Location			0.272
			(0.473)
Constant	8.592***	-0.923	-1.354
	(0.353)	(0.474)	(1.009)
Observations	7,067	7,067	6,804
R-squared	0.007	0.291	0.286

Robust standard errors in parentheses

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

Table 2.16: Linear regression model estimating the effect of income visibility (*revealed* dummy) and log-transformed income on average contribution to the public good. Standard errors clustered on group.

	(1)	(2)	(3)
VARIABLES	Contribution	Contribution	Contribution
1=Revealed	1.745**	-5.229***	-4.790**
	(0.562)	(1.515)	(1.556)
Log(income)		5.416***	5.442***
		(0.463)	(0.464)
Log(income) X Revealed		2.666***	2.462**
		(0.765)	(0.784)
1=Female			-0.362
			(0.507)
Age			0.015
			(0.025)
Location			0.307
			(0.460)
Constant	8.592***	-5.598***	-6.043***
	(0.353)	(0.915)	(1.270)
Observations	7,067	7,067	6,804
R-squared	0.007	0.316	0.310

Robust standard errors in parentheses

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

2.2.3.1.3.2 Revealing incomes reduced inequality

Finally, we assessed the effect of revealing incomes on the level of inequality and the distribution of participant payoffs at the end of the game relative to when incomes were hidden. We computed the Gini index—a commonly used measure of inequality—of the final payoffs of each group. The Gini index is defined as (Allison 1978):

$$G = \frac{1}{2\mu n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}|x_i - x_j|$$

where *n* is the number of players with mean income *μ* over incomes *x*. The Gini takes a value between 0 and 1: the higher the value, the more unequal the set of incomes.

We found that the Gini index at the end of the game is lower in the *revealed* condition (average 0.169) than in the *hidden* condition (average 0.238; Rank-sum, $p < 0.001$). Thus, revealing incomes decreased inequality relative to keeping incomes hidden.

To explain this difference in inequality, we examined contributions over time. To account for multiple testing in these regressions, we report Bonferroni-corrected *p*-values. Participants in quintiles 1 (poorest) through 4 never decreased their contributions in either the *hidden* or *revealed* conditions. In fact, the poorest player in the *hidden* condition actually increased their contributions over time (coeff = 0.047, $p = 0.04$ bonferroni-corrected; all other bottom-to-4th quintiles in both conditions: *p*s = 1.000 corrected; Tables 2.17 and 2.18 cols. 1-4; all regressions were also robust to including demographics).

The contributions of the top quintile participants did, however, differ substantially over time between the *hidden* and *revealed* conditions. In the *revealed* condition, rich participants maintained their contributions over time (coeff = -0.382, $p = 1.000$ corrected, Table 2.18 col. 5), whereas they decreased their contributions over time in the *hidden* condition (coeff = -1.077, $p < 0.001$ corrected, Table 2.17 col. 5). In other words, when incomes were revealed, sanctions were effective in maintaining cooperation from all players, including those with the greatest ability to contribute.

Table 2.17: Linear regression model estimating the effect of round on contribution to the public good in the *hidden* condition. Standard errors clustered on group.

	(1)	(2)	(3)	(4)	(5)
VARIABLES	Quintile 1	Quintile 2	Quintile 3	Quintile 4	Quintile 5
Round	0.047*	0.087	0.092	-0.203	-1.077***
	(0.017)	(0.043)	(0.056)	(0.092)	(0.240)
Constant	2.620***	5.387***	7.549***	10.173***	22.829***
	(0.137)	(0.318)	(0.482)	(0.736)	(1.781)
Observations	683	683	681	682	683
R-squared	0.008	0.006	0.003	0.009	0.039

Robust standard errors in parentheses

*** p<0.001, ** p<0.01, * p<0.05 [Bonferroni corrected]

Table 2.18: Linear regression model estimating the effect of round on contribution to the public good in the *revealed* condition. Standard errors clustered on group.

	(1)	(2)	(3)	(4)	(5)
VARIABLES	Quintile 1	Quintile 2	Quintile 3	Quintile 4	Quintile 5
Round	0.024	0.037	0.032	0.024	-0.382
	(0.021)	(0.061)	(0.073)	(0.103)	(0.266)
Constant	2.663***	5.700***	7.721***	11.385***	25.626***
	(0.157)	(0.369)	(0.434)	(0.587)	(1.723)
Observations	733	733	729	730	730
R-squared	0.002	0.001	0.000	0.000	0.003

Robust standard errors in parentheses

*** p<0.001, ** p<0.01, * p<0.05 [Bonferroni corrected]

2.2.3.2 Experiment 2

Following the procedures of the first experiment, we repeat the same statistical analysis for experiment 2. Unless otherwise noted, all statistics are linear regressions with income quintile as a continuous independent variable and standard errors are clustered on the group. To check the robustness of our results, we also use log-transformed absolute income as continuous IV.

2.2.2.2.1 Individual performance

In experiment 2, participants first completed an individual effort task before they were assigned to groups and received an income level based on their performance in the individual task. Their rank among their group members determined which income level they were assigned to: the best-performing participant in the individual task was allocated the highest income level; the second-best performing participant was assigned the second-highest income level; and so on.

In the individual task, participants had to count the number of 0s in a random matrix of 0s and 1s (Abeler et al. 2011). The more matrices they solved correctly, the higher their performance score. There was no difference in the mean number of correctly solved matrices between the *hidden* (mean = 4.489, s.d. = 1.869) and *revealed* (mean = 4.544, s.d. = 2.029) conditions; $t(438) = -0.298$, $p = 0.766$. There were no significant differences in performance for any quintile between conditions (Table 2.19).

Table 2.19: The number of correctly solved matrices did not differ between conditions for any quintile. Mean values on top, standard deviation in parentheses.

	Hidden condition	*Revealed* condition	Two-tailed *t*-test
Top quintile	6.578 (1.530)	6.953 (1.252)	$p = 0.212$
2nd highest quintile	5.267 (1.074)	5.627 (0.976)	$p = 0.103$
Middle quintile	4.589 (1.125)	4.674 (1.063)	$p = 0.951$
2nd lowest quintile	3.667 (1.000)	3.441 (1.120)	$p = 0.323$
Bottom quintile	2.244 (1.026)	2.023 (1.080)	$p = 0.327$

2.2.2.2.2 Received reward and punishment

In our second experiment, reward and punishment are available in all conditions. Thus, the dependent variable is the number of units that a participant received: negative units represent punishment received, while positive units represent reward received. The independent variable is the income quintile of the recipient of the sanctions (1 to 5). We found qualitatively similar results when we used log-transformed income of the recipient as the independent variable.

In the *hidden* condition, participants could not assess to what extent another player can contribute. Thus, we expected participants to punish the poor for their (low total) contributions and to reward the rich for their (high total) contributions. In the *revealed* condition, conversely, we predicted the mirror image: poorer participants would be rewarded more units for their (high percentage) contribution than wealthier ones.

We estimated the number of units that a participant received as a function of their income quintile and whether the income distribution was hidden or revealed (Table 2.20). In the *hidden* condition, we found that higher income participants received more units (coeff = 0.053, $p = 0.042$, Table 2.20 col. 1), whereas in the *revealed* condition, higher income players received fewer units (coeff = -0.171, $p < 0.001$, Table 2.20 col. 2). Furthermore, a regression including data from both *hidden* and *revealed* conditions together showed that this difference was itself significant (interaction between income and *revealed* dummy, coeff = -0.225, $p < 0.001$, Table 2.20 col. 3).

We found qualitatively similar results with log-transformed income as independent variable: lower income participants received more units in *hidden* (coeff = 0.188, $p = 0.037$, Table 2.21 col. 1) but, conversely, they received fewer units in *revealed* (coeff = -0.582, $p < 0.001$, Table 2.21 col. 2), and the interaction between condition and income was significant (interaction between log-transformed income and *revealed* dummy, coeff = -0.771, $p < 0.001$, Table 2.21 col. 3).

Table 2.20: Linear regression model estimating the effect of a target's income quintile (i.e., their position in the income distribution) on the number of units they received. Standard errors clustered on group.

	(1)	(2)	(3)
VARIABLES	*Hidden*	*Revealed*	Interaction
Quintile	0.054*	-0.171***	0.054*
	(0.026)	(0.030)	(0.025)
1=Revealed			0.949***
			(0.193)
Quintile X Revealed			-0.225***
			(0.039)
Constant	-0.054	0.895***	-0.054
	(0.121)	(0.151)	(0.120)
Observations	1,970	1,935	3,905
R-squared	0.006	0.044	0.043

Robust standard errors in parentheses

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

Table 2.21: Linear regression model estimating the effect of a target's log-transformed income on the number of units they received. Standard errors clustered on group.

	(1)	(2)	(3)
VARIABLES	*Hidden*	*Revealed*	Interaction
Log(income)	0.188*	-0.582***	0.188*
	(0.087)	(0.097)	(0.087)
1=Revealed			1.143***
			(0.217)
Log(income) X Revealed			-0.771***
			(0.130)
Constant	-0.105	1.038***	-0.105
	(0.139)	(0.169)	(0.138)
Observations	1,970	1,935	3,905
R-squared	0.006	0.044	0.043

Robust standard errors in parentheses

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

2.2.2.2.3 Absolute vs. relative contribution

Participants in the *hidden* condition could not take another player's ability to contribute into account since they did not know the income distribution. In the *revealed* condition, on the other hand, participants could evaluate the amount contributed relative to the player's income before choosing whom to punish or reward – in other words, participants could differentiate between absolute and relative contributions.

We hypothesised that absolute contributions would predict the number of units received when income is hidden, but that relative contributions would predict units received when income is revealed. Thus we expected participants to take (give) more units those who give little (a lot) in absolute terms in *hidden*, while their sanctions would be driven by relative contributions in *revealed*.

2.2.2.2.3.1 Absolute vs. relative contribution by quintile

As predicted, higher income participants contribute more in absolute terms in both the *hidden* (coeff = 3.465, $p < 0.001$, Table 2.22 col. 1) and *revealed* (coeff = 5.573, $p < 0.001$, Table 2.22 col. 2) conditions. Conversely, we found that higher income participants contributed a smaller percentage of their income in the *hidden* condition (coeff = -0.064, $p < 0.001$, Table 2.23 col. 1).

Surprisingly, in the *revealed* condition, there was only a weak trend of higher income participants contributing a lower percentage of their income (coeff = -0.0263, $p = 0.108$, Table 2.23 col. 2). This is a slight departure from our previous results in experiment 1: while higher income participants in both experiments contributed more after being punished more and rewarded less in the *revealed* condition, it appears that sanctions were more effective in the *revealed* condition in experiment 2 to encourage richer participants to contribute a higher fraction of their income. These results were qualitatively similar when log-transformed income is used as the independent variable (Tables 2.24 and 2.25).

Table 2.22: Linear regression model estimating the effect of income on absolute contribution. Standard errors clustered on group.

	(1)	(2)
VARIABLES	*Hidden*	*Revealed*
Quintile	3.465***	5.573***
	(0.346)	(0.535)
Constant	-2.446**	-5.640***
	(0.698)	(0.921)
Observations	1,581	1,598
R-squared	0.236	0.381

Robust standard errors in parentheses

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

Table 2.23: Linear regression model estimating the effect of income on percentage of income contributed (relative contribution). Standard errors clustered on group.

	(1)	(2)
VARIABLES	*Hidden*	*Revealed*
Quintile	-0.064***	-0.026
	(0.011)	(0.016)
Constant	0.664***	0.671***
	(0.050)	(0.054)
Observations	1,581	1,598
R-squared	0.055	0.009

Robust standard errors in parentheses

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

Table 2.24: Linear regression model estimating the effect of log-transformed income on absolute contribution. Standard errors clustered on group.

	(1)	(2)
VARIABLES	*Hidden*	*Revealed*
Log(income)	11.742***	18.733***
	(1.151)	(1.761)
Constant	-5.264***	-9.996***
	(0.937)	(1.288)
Observations	1,581	1,598
R-squared	0.235	0.373

Robust standard errors in parentheses

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

Table 2.25: Linear regression model estimating the effect of log-transformed income on percentage of income contributed. Standard errors clustered on group.

VARIABLES	(1) *Hidden*	(2) *Revealed*
Log(income)	-0.210***	-0.084
	(0.037)	(0.054)
Constant	0.710***	0.687***
	(0.058)	(0.065)
Observations	1,581	1,598
R-squared	0.052	0.008

Robust standard errors in parentheses

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

2.2.2.2.3.2 Absolute vs. relative contribution predicted sanctioning behaviour

We hypothesised that participants in the *hidden* condition would reward those who give more in absolute terms (the rich) and punish those who give less in absolute terms (the poor). Conversely, participants in the *revealed* condition would reward those who give a high percentage of their income (poorer participants) but punish those who give a smaller percentage of their income (richer participants).

As predicted, higher absolute contributions in the *hidden* condition led to receiving more reward units and fewer punishment units (coeff = 0.571, p < 0.001, Table 2.26 col. 1). This helped mostly richer participants because they contributed a larger absolute number of units. Relative contributions, in contrast, did not predict the number of units received (p = 0.098, Table 2.26 col. 1) in the *hidden* condition since relative contributions were not observable.

In the *revealed* condition, we found that higher relative contribution led to receiving more units (coeff = -1.775, p < 0.001, Table 2.26 col. 2). Since poorer participants contributed a larger percentage of their income, they were punished less and rewarded more in the *revealed* condition. We also observed that absolute contribution had an effect in the *revealed* condition, in the opposite direction of the effect in the *hidden* condition: higher absolute contributions led to fewer units received (coeff = -0.461, *p* = 0.001, Table 2.26 col. 2).

Table 2.26: Linear regression model estimating the effect of absolute log-transformed contribution and relative contribution on the number of units received in the *hidden* and *revealed* condition. To deal with zero-contributions, a constant of 1 was added to all contributions before applying the log-transformation. Standard errors clustered on group.

	(1)	(2)
VARIABLES	*Hidden*	*Revealed*
Log(contribution+1)	0.571***	-0.461***
	(0.109)	(0.123)
Relative contribution	0.195	1.775***
	(0.115)	(0.212)
Constant	-0.365**	-0.211*
	(0.117)	(0.097)
Observations	1,581	1,598
R-squared	0.084	0.206

Robust standard errors in parentheses

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

2.2.3.1.3.3 Social preference or a desire to take from the rich?

We hypothesised that the poor (rich) would be rewarded (punished) in the *revealed* condition because of their high (low) relative contributions, for which we found evidence presented above. Next we investigated to what extent participants were motivated by more than just relative contributions, such as a desire to reduce the income of the rich regardless of their relative contribution.

Holding constant the fraction of income contributed, we found that richer participants indeed received fewer units (coeff = -0.173, *p* < 0.001, Table 2.27 col. 1). Results were qualitatively similar when we used log-transformed income as the independent variable (coeff = -0.595, *p* < 0.001, Table 2.27 col. 2). Thus, participants gave fewer units to the rich, even when they contributed the same relative amount of their income.

Table 2.27: Linear regression model estimating the effect of relative contribution and income quintile on units received in the *revealed* condition. Standard errors clustered on group.

	(1)	(2)
VARIABLES	Units received	Units received
Relative contribution	1.291***	1.294***
	(0.173)	(0.172)
Quintile	-0.173***	
	(0.032)	
Log(income)		-0.595***
		(0.109)
Constant	0.217	0.367*
	(0.143)	(0.161)
Observations	1,598	1,598
R-squared	0.225	0.226

Robust standard errors in parentheses

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

2.2.2.2.4 Public good provisioning and inequality

2.2.2.2.4.1 Revealing incomes increased contributions

Contributions were higher in the *revealed* than in the *hidden* condition (coeff = 3.134, $p < 0.001$, Table 2.28 col. 1). We observed that higher income participants in the *hidden* condition contributed more than lower income participants (coeff = 3.194, $p = 0.007$, Table 2.28 col. 2); a difference that became significantly larger when incomes were revealed (interaction between income and *revealed* dummy, coeff = 2.109, $p = 0.001$, Table 2.28 col. 2). We found qualitatively equivalent results when using log-transformed income as the independent variable (Table 2.29).

Table 2.28: Linear regression model estimating the effect of income visibility (*revealed* dummy) and income on average contribution to the public good. Standard errors clustered on group.

	(1)	(2)
VARIABLES	Contribution	Contribution
1=Revealed	3.134***	-3.194**
	(0.877)	(1.149)
Quintile		3.465***
		(0.344)
Quintile X Revealed		2.109**
		(0.633)
Constant	7.946***	-2.446***
	(0.467)	(0.695)
Observations	3,179	3,179
R-squared	0.018	0.338

Robust standard errors in parentheses

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

Table 2.29: Linear regression model estimating the effect of income visibility (*revealed* dummy) and log-transformed income on average contribution to the public good. Standard errors clustered on group.

VARIABLES	(1) Contribution	(2) Contribution
1=Revealed	3.134***	-4.732**
	(0.877)	(1.583)
Log(income)		11.742***
		(1.145)
Log(income) X Revealed		6.992**
		(2.092)
Constant	7.946***	-5.264***
	(0.467)	(0.932)
Observations	3,179	3,179
R-squared	0.018	0.333

Robust standard errors in parentheses

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

2.2.2.2.4.2 Revealing incomes reduced inequality

Finally, we assessed the effect that revealing incomes had on inequality. Using (1), we computed the Gini index of the final payoffs of each group. We found that the Gini index at the end of the game was lower in the *revealed* condition (average 0.124) than in the *hidden* condition (average 0.255; Rank-sum, $p < 0.001$). Revealing incomes decreased inequality relative to keeping incomes hidden.

What led to lower inequality in the *revealed* condition? We examined contributions over time. To account for multiple testing in these regressions, we report Bonferroni-corrected *p*-values. Participants in quintiles 1 (poorest) through 4 never decreased their contributions in either the *hidden* or *revealed* conditions (all *p*s > 0.5 corrected; Tables 2.30 and 2.31 cols. 1-4).

Contributions of the highest earners, however, did marginally differ over time between the *hidden* and *revealed* conditions. In the *revealed* condition, rich participants maintained their contributions over time (coeff = -0.056, *p* = 1.000 corrected, Table 2.30 col. 5), whereas they marginally decreased their contributions over time in the *hidden* condition (coeff = -1.105, *p* = 0.080 corrected, Table 2.31 col. 5).

Table 2.30: Linear regression model estimating the effect of round on contribution to the public good in the *hidden* condition. Standard errors clustered on group.

	(1)	(2)	(3)	(4)	(5)
VARIABLES	Quintile 1	Quintile 2	Quintile 3	Quintile 4	Quintile 5
Round	0.029	0.075	0.069	-0.240	-1.105
	(0.021)	(0.067)	(0.203)	(0.190)	(0.442)
Total # rounds	0.036	-0.450*	-0.831***	-0.600	0.957
	(0.105)	(0.144)	(0.211)	(0.302)	(1.591)
Constant	1.287	7.873***	15.421***	15.727***	13.124
	(0.940)	(1.330)	(2.016)	(3.120)	(13.069)
Observations	316	317	316	316	316
R-squared	0.010	0.051	0.061	0.036	0.033

Robust standard errors in parentheses

*** p<0.001, ** p<0.01, * p<0.05 [Bonferroni corrected]

Table 2.31: Linear regression model estimating the effect of round on contribution to the public good in the *revealed* condition. Standard errors clustered on group.

	(1)	(2)	(3)	(4)	(5)
VARIABLES	Quintile 1	Quintile 2	Quintile 3	Quintile 4	Quintile 5
Round	0.025	0.112	0.347	0.192	0.056
	(0.034)	(0.088)	(0.147)	(0.219)	(0.476)
Total # rounds	-0.032	-0.143	-0.360	-0.688	-0.714
	(0.142)	(0.254)	(0.665)	(0.813)	(1.159)
Constant	2.046	5.444*	11.876*	18.442*	31.846**
	(1.266)	(2.354)	(5.782)	(7.564)	(11.144)
Observations	320	320	318	320	320
R-squared	0.003	0.011	0.030	0.012	0.003

Robust standard errors in parentheses

*** p<0.001, ** p<0.01, * p<0.05 [Bonferroni corrected]

Chapter 3.

Preserving the global commons

3.1 Main text

In 1968, Garrett Hardin posed a problem that has remained unsolved in large groups (Hardin 1968). If the success of a public activity depends on voluntary contributions of individuals, then free riders reap larger rewards than contributors and contribution will decline over time. In this seminal and influential paper, Hardin introduced the Public Goods Game, which is a multi-person Prisoner's Dilemma; the latter is focused on two players. While several mechanisms have been described to promote cooperation in pairwise games or in very small groups (Rand and Nowak 2013; Nowak 2006b; Levin 2009), no one has demonstrated a mechanism that allows for the maintenance of cooperation in larger groups. Thus, Hardin's summary, "The population has no technical solution; it requires a fundamental extension of morality," remains unchallenged. Here we propose the first technical solution to this problem.

Experiments focusing on interactions between pairs of people or within small groups (of 3 to 5 people), have established the power of reciprocity for promoting cooperation, be it in the form of repetition (Dal Bo 2005; Fudenberg, Rand, and Dreber 2012), reputation (Wedekind and Milinski 2000; Milinski, Semmann, and Krambeck 2002; Rockenbach and Milinski 2006), shaming (Jacquet 2015; Perez-Truglia and Troiano 2015), network effects (Rand et al. 2014; Fowler and Christakis 2010), threat of expulsion (Cinyabuguma, Page, and Putterman 2005), or costly sanctions (Fehr and Gächter 2000; Gächter, Renner, and Sefton 2008; Rand et al. 2009; Sutter, Haigner, and Kocher 2010; Crockett et al. 2010; Rockenbach and Milinski 2006; Ule et al.

2009). The lever of reciprocity, however, diminishes as group size increases, and therefore these mechanisms, are hard-pressed to promote cooperation on a global scale (Olson 1965; Boyd and Richerson 1988). Although pairs of individuals interacting repeatedly will typically learn to cooperate (Dal Bo 2005), even very small groups interacting repeatedly almost always converge on defection (Grujić et al. 2012).

The reason is that targeted reciprocity is impossible in group interactions: if you stop cooperating, this harms defectors in your group but also cooperators. The problem can be addressed by adding the opportunity for group members to punish or reward each other based on their contributions (Fehr and Gächter 2000; Gächter, Renner, and Sefton 2008; Rand et al. 2009; Sutter, Haigner, and Kocher 2010). Such pairwise interactions allow people to target their reciprocity and have been shown to stabilise cooperation in small groups.

Targeted pairwise interactions, however, cannot scale effectively as groups become larger. With increasing group size, it becomes unlikely that a particular group member has the opportunity to interact with any given other member of the group (Carpenter 2007; Dubreuil 2008). Thus the situation modelled by most previous experiments (Fehr and Gächter 2000; Rand et al. 2009), in which all group members also interact in pairs (and use those pairwise interactions to enforce group cooperation), is untenable when groups are large.

Does this reasoning imply that reciprocity cannot maintain cooperation in large groups? Here we show that the answer is “no.” We demonstrate that coupling a large group cooperative dilemma to a *sparse* network of pairwise reciprocal interactions averts the “tragedy of the commons,” and sustains cooperation in groups an order of magnitude larger than those studied previously. The number of pairwise interactions need *not* scale with the size of the group: a handful of local interactions can support cooperation on a global scale.

To assess the power of such "local-to-global" reciprocity, we developed a novel online software platform called SoPHIE (Software Platform for Human Interaction Experiments, freely available and fully customisable at www.sophielabs.net) to facilitate simultaneous interaction of large numbers of participants (Hendriks 2012). We then used this software to conduct large-scale economic game experiments.

In our first experiment, group sizes were on average 39 people (min = 17, max = 60, sd = 10.28; total N = 646), an order of magnitude larger than typical laboratory experiments with 4 players per group (Fehr and Gächter 2000; Rand et al. 2009). Participants played a repeated 2-stage economic game. In each round of the game, participants first took part in a group contribution stage, and then a pairwise cooperation stage in which they chose actions towards two other group members; for details, see Section 3.1.

In the group contribution stage, participants received an endowment of 20 Monetary Units (MUs), and played a public goods game (PGG) with all other members of the group. In this global interaction, players chose how many of these MUs to contribute to the public good, and how many to keep for themselves. All contributions were doubled and distributed equally among all group members. Thus contributing benefitted the group as a whole, but was individually costly.

For the pairwise cooperation stage, participants were arranged on a ring-structured network in which they were connected to one neighbour on each side (Figure 3.1). Participants played a separate Prisoner's Dilemma (PD) game with each of their two neighbours, who remained the same throughout the experiment. In each PD, participants could cooperate by paying 6 MUs to give the other person 18 MUs, or defect by doing nothing. Participants did not have to take the same action towards both neighbours.

Our experiment had two conditions. In the control condition, local-to-global reciprocity was not possible: in the pairwise cooperation stage, participants were not informed about the group contribution behaviour of their neighbours (Figure 3.1c). Thus they could not use their pairwise relationships to enforce global cooperation, and we expected group contributions to decrease over time.

In the treatment condition, conversely, participants *were* informed of their neighbours' group contributions while making their pairwise cooperation decisions (Figure 3.1d). Thus local-to-global reciprocity was possible, and we expected that (*i*) subjects would preferentially cooperate in the pairwise stage with neighbours that had contributed larger amounts in the group stage; and (*ii*) as a result, we would observe stable high levels of group contribution (in contrast to the control).

To evaluate these predictions, we began by comparing contributions to the group across our two conditions (Figure 3.2a). Indeed, we observed significantly higher average contributions in the treatment compared to the control (coeff = -5.727, $p < 0.001$, Table 3.1; all p-values generated using linear regression with robust standard errors clustered on session, see Section 3.2 for details). Furthermore, this difference in contribution emerged over time: while participants decreased their contributions from round to round in control (coeff = -0.345, $p < 0.001$), contributions in the treatment were stable (no significant decrease in contribution with round, coeff = -0.051, $p = 0.098$; difference between conditions is significant, as shown by the interaction between round and a dummy for the control treatment: coeff = -0.294, $p < 0.001$, Table 3.2).

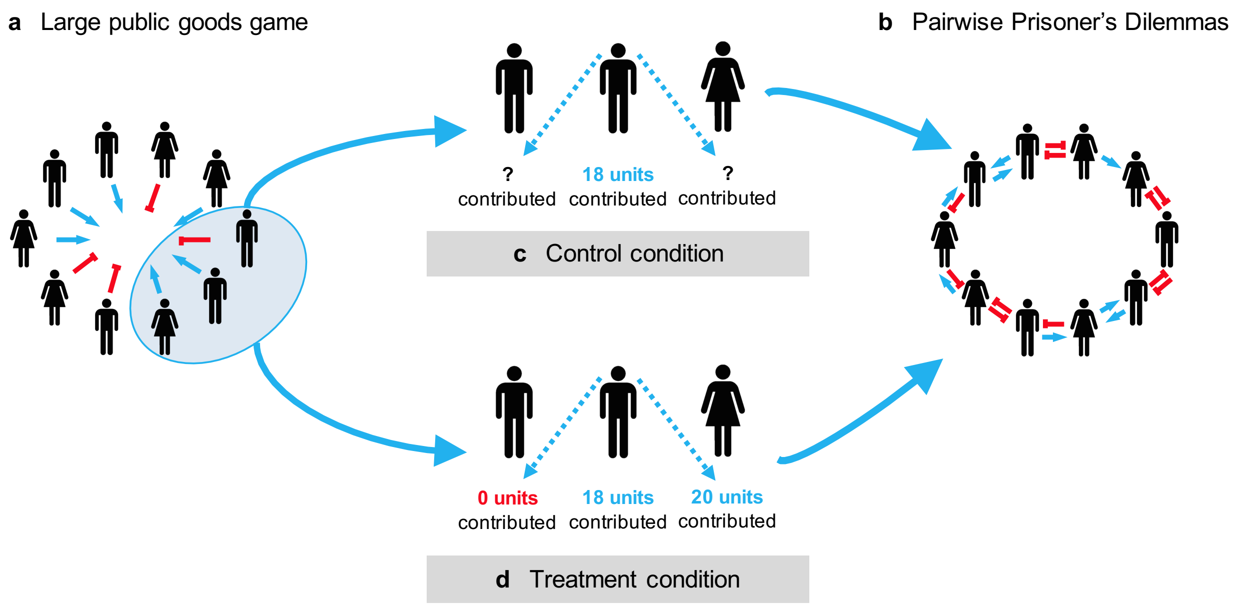

Figure 3.1. The experimental setup consisted of a series of "global" and "local" interactions. ***a,b*** *In each round, participants first took part in a global interaction stage and then in a pairwise local interaction stage.* ***a*** *In the global stage, groups of on average 39 participants (min = 17, max = 60, sd = 10.28, N = 646) played 20 rounds of the Public Goods Game (PGG). In each round, participants were endowed with 20 MUs: they chose how many of these MUs to contribute to a common pool and how many to keep for themselves. The contributed units were doubled and split equally among all group members.* ***b*** *In the pairwise interaction stage, participants were connected to two other group members on a ring-structured network (in experiment 1; for differences to experiment 2, see Section 3.1). In each round, participants played a Prisoner's Dilemma (PD) with each neighbour: they could choose to cooperate by paying 6 units to give 18 units to their neighbours; or defect by doing nothing. Thus mutual cooperation yielded a benefit of 12 for both, unilateral cooperation cost cooperators 6 units while providing defectors with 18 units, and mutual defection did not alter the payoff of either participant.* ***c,d*** *The control and treatment conditions differed in what participants could observe about their neighbours.* ***c*** *In the control condition, participants were not told how many MUs their neighbours contributed in the PGG stage.* ***d*** *In the treatment condition, conversely, participants were informed of their neighbours' contributions in the PGG while making their pairwise decisions in the PD.*

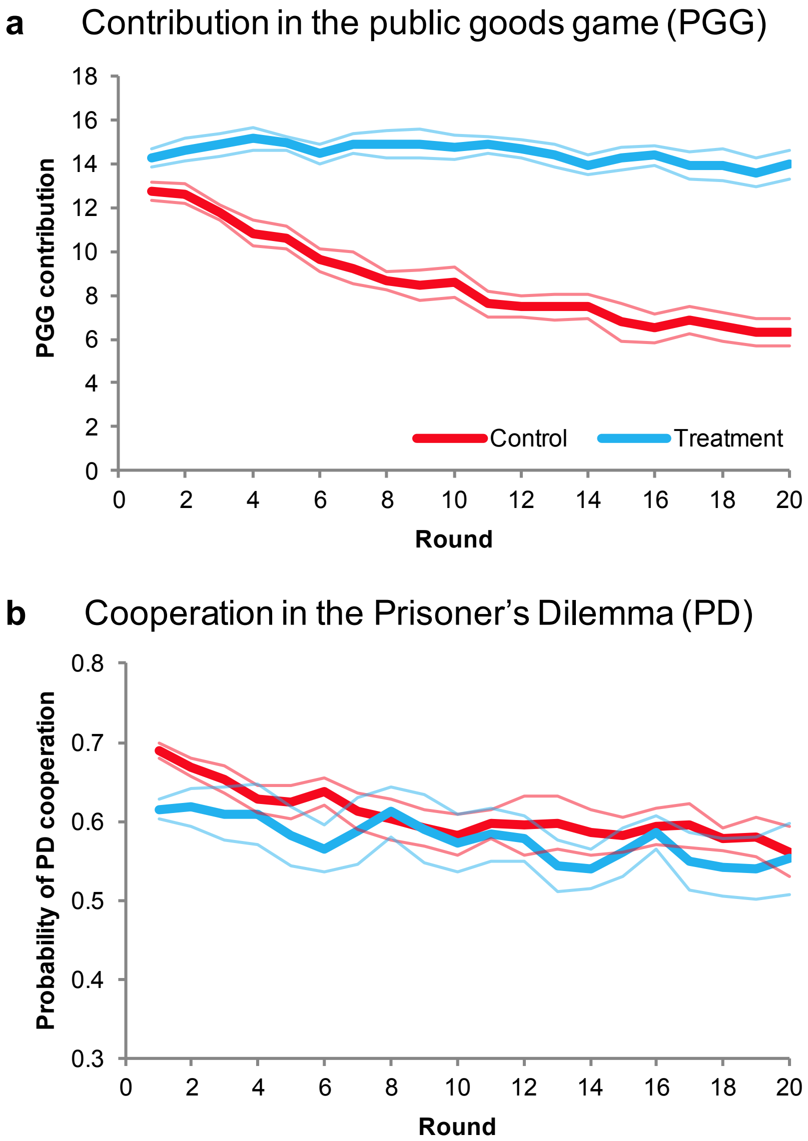

Figure 3.2. Contributions in the PGG were maintained when participants knew their neighbours' previous PGG contributions during the pairwise PD stage. a *PGG contributions were maintained at high levels in the treatment condition when participants were informed of their neighbours' previous PGG contributions. Conversely, in the control condition, the level of contributions in the group cooperation stage decreased quickly over time.* ***b*** *In the pairwise stage, the level of cooperation did not differ between the control and treatment conditions, but the ways in which the pairwise PDs were used differed substantially (see Figure 3.3). (Upper and lower bounds are +/- robust standard errors from the mean clustered on session.)*

What explains the difference in contribution patterns between treatment and control? Participants' pairwise cooperation behaviour provides an answer. While there was no significant difference in *average* levels of PD cooperation between conditions (Figure 3.2b; coeff = 0.031, p = 0.342, Table 3.3), the specific *ways* that PD cooperation was used did differ importantly (Figure 3.3a). In the control, participants were unable to condition their PD cooperation on their neighbours' PGG contributions (since this information was not available). All they could do was cooperate more with a neighbour who cooperated with them in the previous round (coeff = 0.540, p < 0.001, Table 3.5).

In the treatment, on the other hand, participants took advantage of the contribution information available to them to engage in local-to-global reciprocity. In addition to cooperating more with those who previously cooperated with them in the local PD (coeff = 0.475, p < 0.001, Table 3.5), participants were also more likely to cooperate with neighbours who had contributed at least as much as them in the global PGG (coeff = 0.175, p < 0.001, Table 3.5). Moreover, a significant interaction occurred such that participants were most likely to cooperate with neighbours who cooperated in the PD *and* contributed at least as much as them in the PGG (coeff = 0.168, p = 0.002, Table 3.5). Participants in the treatment condition thus reciprocated not only their neighbour's previous pairwise cooperation, but also their contributions in the group cooperation stage: they enacted local-to-global reciprocity. This created an incentive to contribute in the PGG that was absent from the control.

There are two ways that this incentive might be used: did participants reward high contributors by cooperating with them, or punish low contributors by withholding cooperation? To find out, we compared average cooperation rates in control to cooperation rates towards low versus high contributors in treatment (Figure 3.3a). If participants were *rewarding* high contributors, we would expect cooperation rates towards high contributors in the treatment to be higher than the baseline cooperation rate observed in the control. However, we found no such difference (coeff = 0.037, $p = 0.303$, Table 3.7). If, on the other hand, participants were *punishing* low contributors, we would expect less cooperation towards low contributors in the treatment compared to the control baseline; and this is precisely what we observed (coeff = -0.201, $p < 0.001$, Table 3.6). Thus we found evidence that participants in the treatment condition "punished" low contributing neighbours by withholding cooperation.

Finally, we investigated whether this withholding of cooperation from low contributors was effective in eliciting higher PGG contributions in the next round (Figure 3.3b). Interestingly, while receiving PD defection from only one neighbour had no effect on PGG contribution (using number of defecting neighbour as independent variable to predict change in contributions; 1 defecting neighbour: coeff = 0.069, p = 0.871), *both* neighbours defecting in the PD led to a significant increase in PGG contribution in the next round (2 defecting neighbours: coeff = 1.981, p = 0.001, for details see Section 3.2.4). Thus, withholding cooperation was only an effective punishment when both neighbours coordinated their withholding.

In addition to disciplining low contributors, PD cooperation also effectively buttressed high contributors against the temptation to reduce their contributions in treatment: the more PD cooperation high contributors received from their neighbours, the less they reduced their contribution in the next round (coeff = 0.828, $p < 0.001$, Table 3.10). Thus we see a full characterisation of the mechanism by which local cooperation stabilised global contribution.

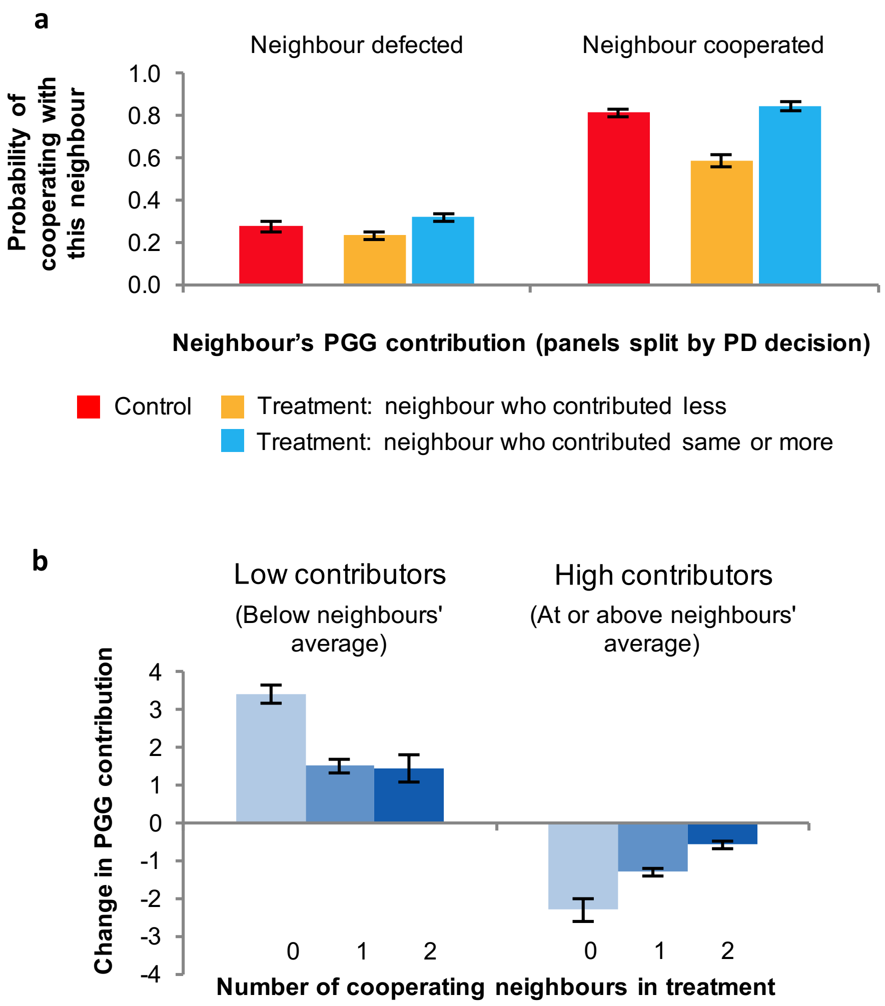

***Figure 3.3. Who is being punished in the pairwise PD stage? a** Participants in the treatment condition were "punished" by not receiving rewards if they had contributed less than their neighbour, compared to the control group. However, participants were not rewarded more than in the control if they contributed at least as much as their neighbour. Thus, local-to-global reciprocity was enacted in local interactions by withholding cooperation from defectors. **b** Participants in the treatment condition respond to their neighbours' decision to cooperate or defect in the pairwise cooperation stage: when both neighbours withheld cooperation from participants who contributed less in the PGG than their neighbours, participants increased their contributions in the PGG in the subsequent round. Conversely, local-to-global reciprocity was buttressing against the temptation to defect: the more PD cooperation high-contributing participants received from their neighbours, the less they decreased their contributions. (Error bars represent robust standard errors clustered on session.)*

Importantly, these effects were unique to treatment: participant in the control condition did not change their contribution behaviour in response to amount of PD cooperation they received (low contributors: coeff = -0.048, *p* = 0.857, Table 3.8; high contributors: coeff = 0.253, *p* = 0.183, Table 3.9), and this differed significantly from what we observed in the treatment (interaction between number of cooperating neighbours and control dummy; low contributors: coeff = 1.105, *p* = 0.002, Table 3.8; high contributors: coeff = -0.575, *p* = 0.015, Table 3.9).

Finally, we present evidence that the power of local-to-global reciprocity is *scalable*. First, we take advantage of random variation across sessions in the number of participants in the PGG. One might worry that as groups become larger, local interactions with just two neighbours would become less effective at maintaining global cooperation. However, we find no evidence of this: a threefold increase in the size of the group had no discernible impact on PGG contributions in the treatment (using group size of each session as independent variable to predict the average contribution in the final round of the game in treatment: coeff = -0.015, *p* = 0.782, Table 3.11). Our intervention was just as effective relative to the control (or perhaps even *more* effective) for preserving the global commons in large groups compared to small groups (Figure 3.4a).

To further demonstrate the scalability of our intervention, we conducted a second experiment with a *much* larger PGG group of 1000 people. Participants in the second experiment played a repeated two-stage economic game that was identical to the first experiment, except that in the pairwise cooperation stage, participants played a PD with just one other member of the group (rather than two others, as in the first experiment). We reduced the number of PD partners to further assess the robustness of our “local-to-global” intervention; for details on the experimental design and differences to the first experiment, see Section 3.1.

Despite the extremely large group size of our second experiment, we replicated our earlier results. Average contributions were significantly higher in treatment than in control (coeff = 1.456, *p* = 0.005, Table 3.13), and this difference emerged over time (interaction between control dummy and round number, coeff = -0.1092, *p* = 0.017, Table 3.13): while contributions in the treatment were stable (coeff = -0.027, *p* = 0.429), contributions in the control condition decreased with round (coeff = -0.136, *p* < 0.001) (Figure 3.4b).

In summary, we have shown that "local-to-global" reciprocity can maintain stable contributions in a large public goods game. Participants punished other group members who contributed less than them by withholding cooperation. Low contributors, in turn, increased their contributions when their neighbours jointly withheld cooperation from them, while high contributors continued to contribute when their neighbours cooperated with them. Thus, stable levels of contributions emerged in the group cooperation stage of the treatment. In the control, conversely, such local-to-global reciprocity was not possible, and PGG contributions collapsed.

Across two experiment, we found that group size did not affect our results: contributions in treatment were sustained in groups several magnitudes larger than previously studied. Thus, we have demonstrated that targeted reciprocity need not be scaled with the size of the network: instead participants only need to be informed of what a small number of other participants in the network who they interact with did previously.

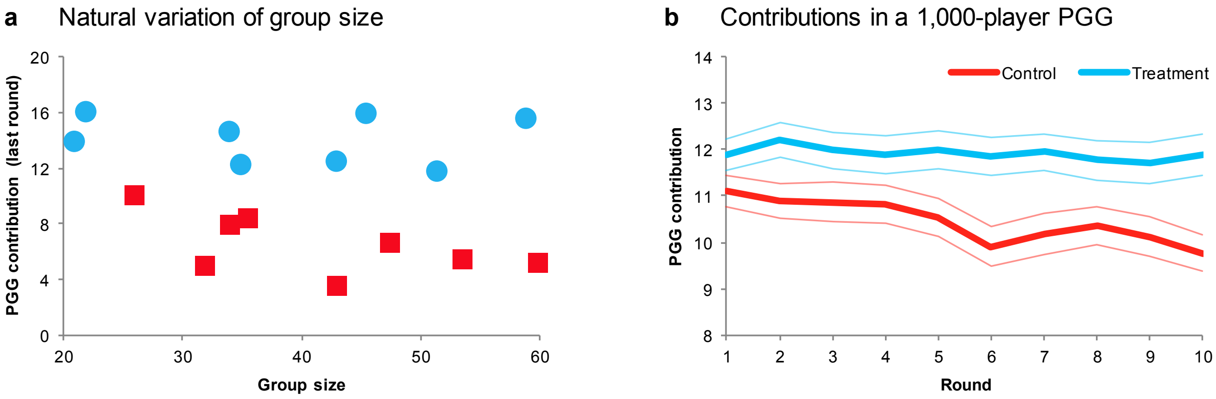

***Figure 3.4. "Local-to-global" reciprocity is invariant to the size of the group. a** We take advantage of random variation across sessions in the number of participants: the size of the group does not have an effect on the level of contributions in the final round of the game in the treatment condition. Indeed, a threefold increase in group size does not affect contributions when "local-to-global" reciprocity is possible. **b** In a second experiment (see Section 3.1), we recruited 1,000 participants to play the same large-scale PGG over 10 rounds. Participants in treatment, where "local-to-global" reciprocity with the PD partner was possible, made stable PGG contributions, while participants in control decreased their contributions over time. (Upper and lower bounds are +/- robust standard errors clustered on double-pairs; see Section 3.2 for statistical details.)*

Theorists have argued that group size poses a challenge for reciprocity-based mechanisms in sustaining cooperation and cannot readily explain the levels of cooperation observed in contemporary and ancient societies (Olson 1965; Boyd and Richerson 1988; Esteban and Ray 2001). But these theories did not consider the possibility of pairwise interactions that allow for targeted action. Our findings show that direct reciprocity *can* in fact maintain cooperation in large groups, if each individual has even a very small number of pairwise interactions. Developing theoretical models of the interaction between pairwise and group interactions is an important direction for future research.

Our findings build on existing interventions to increase public goods contributions in the real world that have implications for policy-makers (Kraft-Todd et al. 2015; Weber and Johnson 2012). Sign-ups among residents in apartment complexes to participate in a voluntary energy reduction program are higher when the sign-up sheet is publicly observable (Yoeli et al. 2013). The more tax evaders are aware that their neighbours know of their delinquency, the higher their compliance with tax repayments (Perez-Truglia and Troiano 2015). Our laboratory experiments provide tightly controlled evidence of the mechanism underpinning these field experiment results: when we are provided with information about other people's cooperative actions, we will reward them for their contributions to our community and to the world at large, allowing us to finally answer Hardin's challenge to preserve the global commons.

3.2 Supporting figures and data

3.2.1. Methods

3.2.1.1 Data collection

We recruited U.S. participants for both experiments from the online labour market Amazon Mechanical Turk (AMT). AMT is an online market place in which employers can pay users for completing short tasks – usually referred to as Human Intelligence Tasks (HITs) – for a relatively small pay (generally about $1.00 for 10 minutes of work).

AMT has been shown to be more diverse and more nationally representative than the typical college student sample at major research universities (Rand, Greene, and Nowak 2012; Amir et al. 2012; Horton, Rand, and Zeckhauser 2011). Workers who have been recruited on AMT receive a baseline payment and can also be paid a bonus depending on their performance in the task. This setup lends itself well to adopt incentivised economic experiments: the baseline payment acts as the 'show-up' fee and the bonus payment may derive from the workers' behaviour in the economic game and/or other tasks throughout the experiment.

There may, of course, exist potential issues on AMT that would not occur in a traditional laboratory setting. For instance, running an experiment online involves giving up some control over subjects, since they cannot be monitored, as is usually the case in laboratories. That is, it cannot be ruled out that more than a single person is taking part in the experiment or that one person is participating more than once in the experiment (although AMT has put extensive measures into place to avoid this from happening; in addition, we have also implemented ways to carefully screen out any possible re-takers). Finally, the participating subject sample, albeit more diverse and representative than the average college students sample, is biased towards those who participate in online labour markets in the first place. To address these possible concerns, numerous studies have been carried out to validate results collected using AMT (Amir et al. 2012; Horton, Rand, and Zeckhauser 2011; Berinsky, Huber, and Lenz 2012).

Our experiments (described in detail below) were implemented using the interactive experimental platform SoPHIE (Software Platform for Human Interaction Experiments), which is freely available and fully customisable at www.sophielabs.net (Hendriks 2012).

In experiment 1, we recruited a total of 646 participants across 16 sessions. Each session lasted for approximately 35-40 minutes. All participants who completed the experiment earned a $3.00 show-up fee and had the opportunity to earn an additional “bonus” payment depending on their and others’ decisions in the public goods game and the prisoner’s dilemma. Average earnings from the game including bonus were $4.34.

In experiment 2, a total of 1,352 participants were recruited across 15 sessions. Each session lasted for approximately 15-20 minutes. All participants who completed the experiment earned a $1.00 show-up and could earn a “bonus” payment depending on their decisions and those of other participants in the experiment. Average earnings from the game including bonus were $1.34.

All experiments were approved by Harvard University Committee on the Use of Human Subjects in Research.

3.2.1.2 Experimental design

3.2.1.2.1 Experiment 1

Participants on AMT joined the experiment by responding to our ‘HIT’ posted on the AMT website, and being redirected to our external website where the game was hosted. They then received instructions on the experimental game and had to pass a comprehension quiz about the rules of the game (see Section 3.3.1 for screenshots of instructions and comprehension questions). Participants were not allowed to continue unless they answered all three comprehension questions correctly. Participants were then asked to wait up to 10 minutes for other participants to arrive before the experiment began; once the experiment did begin, all participants started at the same time. A countdown was displayed on their screen for the last three minutes and an audio feedback was played informing them about the remaining time until the experiment would start. Participants who did not respond within 40 seconds after the start of the experiment could not participate in the experiment. All participants were informed upfront that their presence was mandatory to be eligible to take part in this study. AMT workers who had taken part in a previous session of this experiment were not allowed to participate again.

The experiment was conducted one session at a time. For each session, we aimed to maximise the number of participants. The average group size was 39 participants (min = 17, max = 60, sd = 10.28). We launched our experiment only during business hours (9am – 5pm Eastern Standard Time) on weekdays for every session. All participants were assigned to the same condition during a single session. We randomized the order of treatment and control conditions across sessions (8 treatment and 8 control) prior to the start of the experiment.

All participants who were eligible to play (i.e., finished the instructions and the quiz in the allotted time) were arranged in a circular network so that every participant had exactly two neighbours (see Figure 3.1 of the main text). The network structure did not change over the course of the experiment, except as noted below.

Prior to the beginning of the actual game, all participants took part in a practice round. The practice round was played with two neighbours simulated by a computer, which participants were informed about. The practice round took place simultaneously for all participants to ensure that all participants were paying attention and were ready for the actual game. (See Section 3.3 for screenshots of the practice rounds.) During the practice round, all times to reach a decision were doubled, from 20 seconds to 40 seconds, to ease familiarisation with the setup.

After the practice round, the real game began with participants interacting with their two neighbours. Due to a technical error, in all sessions neighbours were randomly reassigned after round 1 (but were not informed of this reshuffling). We believe this error is unlikely to have had any long-lasting consequences for our participants; and whatever consequences it might have had would have worked against our treatment effect, by undermining the power of local-to-global reciprocity after the first round.

From the second round onwards, a participant's neighbours stayed the same as long as neither the participant nor her neighbours dropped out of the game (participant dropouts are a common problem in online studies, unlike in the physical lab, and the solution we take here is standard procedure, see Ref. (Rand, Arbesman, and Christakis 2011; Rand et al. 2014)). Dropouts were eliminated from the circular network and the dropouts' former neighbours were connected. Participants were not told if their neighbour dropped out to avoid a 'restart' effect which has been observed in repeated games (Andreoni 1988; RTA Croson 1996). Participants were told to pay full attention and to avoid dropping out, or else their payoff—show-up fee and bonus—would be zero.

Since dropouts did occur, one might worry about potential selection effects. Most importantly, there was no difference between the treatment and control in dropout rate (logistic regression using treatment dummy to predict probability of dropout, standard errors clustered on session, $p = 0.752$) or average group size (t-test of group size between conditions using a single indicator variable per condition, $p = 0.690$). Thus, differences in behaviour between the treatment and control cannot be attributed to dropouts. Furthermore, we did not find evidence that the behaviour of dropouts was systematically different from non-dropouts: there was no statistical difference in contributions between dropouts and non-dropouts (linear regression using dropout dummy to predict contributions clustered on session, $p = 0.144$), and contribution amount did not predict the probability of dropping out (logistic regression using contributions to predict likelihood of dropout clustered on session, $p = 0.121$).

The experiment consisted of a series of 20 rounds. Participants were not told how many rounds they would be playing to avoid potential last-round effects and backwards induction (as in Ref. (Rand et al. 2009)). Each round was comprised of a public goods game (PGG; stage 1) with all participants in the session contributing to a shared pool, followed by pairwise Prisoner's Dilemmas (PDs; stage 2) between the direct neighbours in the circular network.

In stage 1, participants chose a contribution of between 0 and 20 units in the PGG. All contributions were doubled and every participant in the session received an equal share from the public good. After making their PGG contribution decision, participants in both conditions learned their individual payoff from the PGG. In the treatment condition, the participants were also informed of their neighbours' contributions to the PGG, while participants in the control condition received no additional information.

Across both conditions, participants in stage 2 then played two pairwise PDs with their two neighbours. They could choose between cooperation (paying a cost of 6 units to provide the neighbour with a benefit of 18 units) and defection (paying no cost and providing no benefit). Once all participants had made their choice, in both conditions the PD actions of the participant's two neighbours were displayed and the participant's payoffs in the current round were summarised. (See Section 3.3 for instructions and screenshots of the experiment.)

3.2.1.2.2 Experiment 2

Participants on AMT joined the experiment by accepting our AMT ‘HIT’. They read the instructions (see screenshots in Section 3.3) and had to pass several comprehension questions. Participants waited in an online ‘waiting room’ for up to 5 minutes for three other participants to arrive. As soon as four participants were ready, the game began immediately. There was no practice round; however, the time to reach a decision in each stage was 10 seconds longer in the first round of the experiment than in later rounds.

The two-stage economic game in the second experiment was similar in many ways to one in the first experiment: participants first made a decision in a group contribution stage—the large-scale repeated PGG—and then in a pairwise cooperation stage—a repeated PD. However, there were also several differences between the two experiments. First, while participants in the first experiment interacted with two players in the PD stage, every participant in experiment 2 played a repeated PD with only one other participant.

Furthermore, participants in this experiment were no longer recruited all at the same time; they were instead recruited in batches of 4 participants, which formed two pairs who played the game at the same time. (We refer to these two simultaneously playing pairs of players as “double pairs.”) We required two pairs of participants playing the game simultaneously due to a change to the control condition. Participants in the control in experiment 1 were not informed about anyone’s PGG contributions during the PD stage, while participants in the treatment condition always knew the PGG contributions of two other players (those with whom they also played the PD). We argued in experiment 1 that observing the PGG contributions of one’s PD partners is crucial to sustaining contributions; however, an alternative “social norm” explanation could be that seeing *anyone else’s* PGG contributions is sufficient to maintain contributions (i.e., without playing a repeated

PD with that same person). We rule out this "social norm" possibility with a new control condition in experiment 2.

Participants in the control condition in experiment 2 saw the PGG contributions of another player who was part of the larger PGG. This player was *not* the same participant with whom they interacted in the PD stage of the game. They saw instead the contributions of one of the players of the pair that played the game simultaneously, and played a repeated PD with a participant whose contributions they did not see. In the treatment condition, conversely, participants continued to observe the contributions of the player with whom they also played the PD game. Thus, in both conditions, participants saw *someone's* PGG contributions: any difference we observe between control and treatment thus cannot be attributed to a "social norm" of others' contribution, but is caused by interacting directly with the person whose contributions were observable. (While we required four participants to be playing the game at the same time in the control, we only needed two players at the same time in the treatment condition. To avoid any differences in decision times or dropout rates between conditions, however, participants in the treatment condition also played the game in batches of 4 participants. Note though that each pair of players played their game independently and was not aware of another pair that played simultaneously.)

Finally, participants in the second experiment did not learn about their payoff from the PGG. Because all 1,000 participants were not online simultaneously, it was not possible to calculate the payoff of each round of the PGG in real time. Participants were told that their earnings from the PGG would be calculated at the end of the study. Thus, participants in neither condition learned whether or not overall levels of contributions in the large group were stable, decreasing, or increasing. The lack of feedback from the PGG implies that conditionally cooperative players (Fischbacher, Gächter, and Fehr 2001) would not be able to respond to the changes in contributions by the entire group. However, they would still be able to observe, and respond to, the PGG contributions of the player whose contributions they saw during the PD stage. Furthermore, although the lack of PGG feedback could potentially affect individuals' contribution behaviour, this lack of feedback is the same across both conditions and it could thus not drive any difference between conditions.

The experiment consisted of a series of 10 rounds of the two-stage economic game (as described in more detail above for experiment 1). To avoid end-game effects, participants were not told how many rounds would be played (e.g. see (Rand et al. 2014)). In stage 1, participants could contribute between 0 and 20 units in the PGG. In stage 2, participants played a PD with another player who remained the same throughout the game: each person could choose to cooperate (paying 12 units to increase the other player's payoff by 36 units) or defect (no cost or benefit to either party) with the other player.

The experiment was conducted one session at a time. For each session, we recruited as many as 200 participants per session, and recruitment continued until we had 500 participants per condition (total N = 1,000) who had completed the game. To keep with random assignment, the order of conditions was alternated across sessions. All participants were assigned to the same condition for every session. In total, we conducted 15 sessions (7 control, 8 treatment) and recruited 1,352 participants, of which 26% of groups did not complete the game due to one or more dropout.

Dropouts in the second experiment were handled differently than in the first experiment. While in experiment 1 a participant who dropped out was simply "replaced" by his or her two nearest neighbours joining the cyclic network and playing the remainder of the game together, this was not possible in experiment 2, as there were only 4 participants in the same stage of the game at the same time. Thus, if one participant dropped out (e.g., by closing his or her Internet browser, or losing Internet connection), the remaining three participants, who were part of the two pairs playing the game simultaneously, could not continue. Although those participants could not finish the game, they were compensated for their time by earning the $1.00 show-up fee and a bonus of $0.30.

Across both conditions, 352 participants (26%) did not complete the game. There was no significant difference in the number of dropout groups between conditions (logistic regression using treatment dummy to predict probability of dropout at the "double pairs" level, with robust standard errors, $p = 0.714$). Neither did levels of contributions nor rates of pairwise cooperation predict the probability of dropping out (logistic regression to predict probability of dropout, clustered on double pair; using contribution: $p = 0.473$; using cooperation: $p = 0.378$).

Our main analysis focuses on the 1,000 participants (500 per condition) who completed all 10 rounds of the game. However, we find qualitative similar results when dropout groups are included (see Table 3.13).

3.2.2. Statistical details

In experiment 1, all games lasted 20 rounds. In each round of the game, participants had to make three choices. First, how many units they wanted to contribute to a group-wide PGG. Contributions in the PGG are measured on a continuous scale (i.e., integers from 0 to 20 where 0 is full defection and 20 is full cooperation). Then, they made two simultaneous decisions in the PD stage: whether or not to cooperate with each of their two neighbours. Cooperation in the PD is a binary measure (i.e., 1=cooperation, 0=defection).

In experiment 2, all games lasted 10 rounds. In each round of the game, participants made two choices: how many units (between 0 and 20) to contribute in a large-scale PGG and whether or not to cooperate with another participant in the PD.

Unless otherwise indicated, we used linear regression models with robust standard errors clustered on session to account for the fact that decisions of players within a given session are not independent.

3.2.2.1 Group contributions

We first asked the basic question of how contributions in the group cooperation stage differed between conditions. We predicted contributions to the public good in the control condition to be less than in the treatment condition. This difference would grow as time passed: participants would maintain stable contributions in the treatment condition, while contributions in the control condition would decrease over time. The dependent variable in our analyses was the amount of units contributed per round. The independent variables were a dummy for the control condition (1=control, 0=treatment) and current round number.

As predicted, we found that participants contributed less on average in the control condition than in the treatment condition, both in the first round (coeff = -1.491, $p = 0.018$, Table 3.1 col. 1) and averaged over all rounds (coeff = -5.727, $p < 0.001$, Table 3.1 col. 2). Furthermore, this difference in contributions emerged over time (interaction between round and control dummy, coeff = -0.294, $p < 0.001$, Table 3.2 col. 3): we observed a significant decrease in contribution over time in the control (coeff = -0.345, $p < 0.001$, Table 3.2 col. 1), but not in the treatment (coeff = -0.051, $p = 0.098$, Table 3.2 col. 2).

Table 3.1: Linear regression model estimating the effect of treatment on contributions in the group cooperation stage. The treatment condition is taken as baseline. Standard errors clustered on session.

	First round	All rounds
1=Control	-1.491	-5.727
	(0.561)*	(0.633)***
Constant	14.231	14.484
	(0.413)***	(0.409)***
R^2	0.01	0.12
N	646	11,552

* *p*<0.05; ** *p*<0.01; *** *p*<0.001

Table 3.2: Linear regression model estimating the effect of round and experimental condition on contributions in the group cooperation stage. In column 3, the treatment condition is taken as baseline. Standard errors clustered on session.

	Control	Treatment	Both
Round	-0.345	-0.051	-0.051
	(0.026)***	(0.027)	(0.026)
1=Control			-2.763
			(0.546)***
1=Control X round			-0.294
			(0.036)***
Constant	12.240	15.003	15.003
	(0.460)***	(0.328)***	(0.317)***
R^2	0.06	0.00	0.15
N	5,981	5,571	11,552

* *p*<0.05; ** *p*<0.01; *** *p*<0.001

3.2.2.2 Pairwise cooperation

We then turned to the question of how participants interacted in the pairwise cooperation stage. Participants could choose to cooperate or defect with each of their neighbours (They did not have to make the same choices for both.)

Here, our unit of observation was the PD cooperation decision (2 observations per participant per round). The independent variable was PD choice (0=defect, 1=cooperate). The dependent variables were a dummy for the control condition (1=control, 0=treatment) and current round number. We use linear regression (despite having a binary DV) in order to have more easily interpretable coefficients; however, we note that using logistic regression instead does not qualitatively change any outcomes.

Although we found that there was significantly more cooperation in the control condition than the treatment in period 1 (coeff = 0.075, p < 0.001, Table 3.3 col. 1), there was no significant difference when considering all rounds (coeff = 0.031, p = 0.342, Table 3.3 col. 2). Furthermore, there was no significant difference between conditions in how cooperation changed over time (interaction between round number and control dummy, coeff = -0.001, p = 0.492, Table 3.4 col. 3): cooperation declined very slightly over time in both the control condition (coeff = -0.005, p = 0.016, Table 3.4 col. 1) and treatment condition (coeff = -0.004, p = 0.006, Table 3.4 col. 2), at a modest rate of on average 0.4% per round.

Table 3.3: Linear regression model estimating the effect of treatment on levels of cooperation in the pairwise cooperation stage. The treatment condition is taken as baseline. Standard errors clustered on session.

	First round	All rounds
1=Control	0.075	0.031
	(0.015)***	(0.032)
Constant	0.615	0.579
	(0.012)***	(0.028)***
R^2	0.01	0.00
N	1,292	23,104

* *p*<0.05; ** *p*<0.01; *** *p*<0.001

Table 3.4: Linear regression model estimating the effect of round and experimental condition on PD cooperation. In column 3, the treatment condition is taken as baseline. Standard errors clustered on session.

	Control	Treatment	Both
Round	-0.005	-0.004	-0.004
	(0.002)*	(0.001)**	(0.001)**
1=Control			0.044
			(0.027)
1=Control X round			-0.001
			(0.002)
Constant	0.661	0.617	0.617
	(0.012)***	(0.026)***	(0.025)***
R^2	0.00	0.00	0.00
N	11,962	11,142	23,104

* *p*<0.05; ** *p*<0.01; *** *p*<0.001

3.2.2.3 Pairwise cooperation strategies

While the average levels of cooperation in the PD stage did not differ between the two conditions, the *ways* in which the PD was used did differ between conditions.

In the control condition, participants could not condition their behaviour in the PD on their neighbours' PGG contributions, because this information was not available to them. Thus, we only expected participants to condition their PD behaviour on their neighbours' previous cooperation (i.e. to engage in "local" reciprocity). Indeed, we found that participants in the control condition were substantially more likely to cooperate if their partner had cooperated with them in the previous round (using neighbour's action in prior PD round as independent variable with 0=defect, 1=cooperate: coeff = 0.540, p < 0.001, Table 3.5 col. 1).

Conversely, participants in the treatment condition were informed of their neighbours' PGG contributions while making their PD decisions. They were thus able to enact local-to-global reciprocity: they could condition their local PD cooperation with a given neighbour on that neighbour's contribution to the global PGG. Indeed, participants in the treatment were significantly more likely to cooperate with neighbours who were high contributors in the PGG (using neighbour's action in PGG immediately prior to the given PD as independent variable with 0=neighbour contributed less than the participant, 1=neighbour contributed at least as much as the participant, following the definition in (Rand et al. 2009): coeff = 0.175, p < 0.001, Table 3.5 col. 2).

Participants also engaged in traditional local reciprocity, cooperating more with neighbours who had cooperated with them in the previous PD round (coeff = 0.475, p < 0.001, Table 3.5 col. 2). Furthermore, there was a synergistic interaction between local reciprocity and local-to-global reciprocity (interaction between neighbour's cooperation dummy and neighbour's contribution dummy, coeff = 0.168, p = 0.002; Table 3.5 col. 3), such that participants were most likely to cooperate with neighbours who both cooperated in the previous PD *and* were high contributors in the PGG.

Table 3.5: Linear regression model estimating the effect of a neighbour's previous pairwise cooperation and her group contribution on the participant's willingness to cooperate with her in the current round. Standard errors clustered on session.

	Control	Treatment	Treatment
1=Neighbour cooperated	0.540 (0.038)***	0.475 (0.014)***	0.356 (0.033)***
1=Neighbour contributed same or more than me		0.175 (0.021)***	0.088 (0.027)*
1=Neighbour cooperated X contributed same or more			0.168 (0.035)**
Constant	0.274 (0.026)***	0.176 (0.012)***	0.233 (0.017)***
R^2	0.29	0.27	0.28
N	11,294	10,518	10,518

* p<0.05; ** p<0.01; *** p<0.001

Two mechanisms could explain our results: either participants cooperated in the PD more with high-contributing neighbours in the PGG, or they punished low-contributing neighbours by withholding cooperation from them in the PD (or both). To find out which mechanism was at work in our data, we compared cooperation rates towards low and high contributors in the treatment condition to cooperation rates in the control condition (where the neighbour's contribution was unknown).

If punishment of low contributors was occurring, we would expect less PD cooperation with low contributors in the treatment than with unknown contributors in the control; and indeed, this is what we observe (regression including all PD choices from control and PD choices where neighbour contributed less than the participant from the treatment, using treatment dummy as independent variable: coeff = -0.201, p < 0.001, Table 3.6 col. 1; controlling for previous cooperation behaviour does not affect this result, coeff = -0.135, p < 0.001, Table 3.6 col. 2). If reward of high contributors was occurring, conversely, we would expect more PD cooperation with high contributors in the treatment than with unknown contributors in the control; but we find no such effect (regression including all PD choices from control and PD choices where neighbour contributed as much as more than the participant from the treatment, using treatment dummy as independent variable: coeff = 0.037, p = 0.303, Table 3.7).

In short, pairwise cooperation rates in the treatment condition differed towards low-contributing neighbours relative to the control group, but not towards high contributors. Thus participants "punished" low contributors by withholding cooperation, rather than rewarding high contributors by increasing cooperation. This suggests a norm of where people are expected to contribute, such that deviations downward are punished.

Table 3.6: Linear regression model estimating the effects of the treatment on a participant's likelihood of cooperating in the pairwise cooperation stage with a neighbour who contributed *less than* the participant in the group cooperation stage. The baseline group are participants of any contribution level in the control condition. Standard errors clustered on session.

	Neighbour contributed less	Neighbour contributed less
1=Treatment	-0.201 (0.026)***	-0.135 (0.018)***
1=Neighbour cooperated		0.500 (0.034)***
Constant	0.610 (0.016)***	0.299 (0.024)***
R^2	0.03	0.27
N	15,141	14,270

* p<0.05; ** p<0.01; *** p<0.001

Table 3.7: Linear regression model estimating the effects of the treatment on a participant's likelihood of cooperating in the pairwise cooperation stage with a neighbour who contributed the *same or more than* the participant in the group cooperation stage. The baseline group are participants of any contribution level in the control condition. Standard errors clustered on session.

	Neighbour contributed same or more	Neighbour contributed same or more
1=Treatment	0.037 (0.034)	0.036 (0.019)
1=Neighbour cooperated		0.534 (0.023)***
Constant	0.610 (0.016)***	0.278 (0.017)***
R^2	0.00	0.29
N	19,925	18,836

* p<0.05; ** p<0.01; *** p<0.001

3.2.2.4 Response in PGG to neighbours' PD choices

We observed that participants in the treatment condition were less likely to cooperate in the PD stage with neighbours who had contributed less than them in the PGG. They withheld cooperation to "punish" low contributors. Did this withholding work to elicit more contributions from low contributors in the future?

Indeed, we found that in the treatment, the less a low contributor's neighbours cooperated with her, the more she contributed in the next PGG round (change in contribution predicted by the number of neighbours withholding cooperation (i.e. defecting in PD): coeff = 1.153, p < 0.001, Table 3.8 col. 2). Interestingly, withholding had to be coordinated in order to be effective: having only one neighbour withhold cooperation did not increase subsequent contributions of the low contributor relative to having both neighbours cooperate (coeff = 0.069, p = 0.871); it was necessary to have *both* neighbours withhold cooperation in order to motivate low contributors to increase their contributions (0 vs. 2 withholding neighbours: coeff = 1.981, *p* = 0.001) (see Table 3.9).

Importantly, this effect was unique to the treatment. In the control, having cooperation withheld by one or both neighbours had no effect on low-contributing participants' subsequent PGG contribution (using number of withholding neighbours as independent variable: coeff = 0.048, *p* = 0.857, Table 3.8 col. 1; using discrete number of neighbours withholding: 0 vs 1 neighbour withholding: coeff = -0.038, *p* = 0.920; 0 vs 2 cooperating withholding: coeff = 0.129, *p* = 0.804, Table 3.9 col. 1). Furthermore, when data from both conditions are taken together, a significant interaction between condition and the number of neighbours withholding cooperation demonstrated that the effect of withholding on future contributions was significantly larger in the treatment than the control (interaction between number of cooperating neighbours and treatment dummy: coeff = 1.105, *p* = 0.002, Table 3.8 col. 3; qualitatively similar results as above using interaction between treatment dummy and discrete number of neighbours, see Table 3.9 col. 3).

In addition to disciplining low contributors, neighbours' behaviour in PD mechanism also effectively buttressed high contributors against the temptation to reduce contributions in the treatment condition: the more PD cooperation high contributors received from their neighbours, the less they reduced their contributions in the next round (coeff = 0.828, *p* < 0.001, Table 3.10 col. 2). In the control condition, however, there was no "buttressing effect": receiving more cooperation from neighbours did not protect against declining contributions in control (coeff = 0.253, *p* = 0.183, Table 3.10 col. 1); an observation that was also confirmed by a significant interaction for the treatment condition only (interacting number of cooperating neighbours with treatment dummy: coeff = 0.575, *p* = 0.015, Table 3.10 col. 3).

Table 3.8: Linear regression model estimating the effect of both neighbours' defection in the PD stage on change in contributions of participants who contributed *less than* their neighbours previously in the PGG stage. Standard errors clustered on session.

	Control	Treatment	Both
# Neighbours withholding PD cooperation	0.048 (0.255)	1.153 (0.183)***	0.048 (0.246)
1=Treatment			-0.966 (0.369)*
1=Treatment X neighbours withholding			1.105 (0.303)**
Constant	1.840 (0.247)***	0.874 (0.291)*	1.840 (0.239)***
R^2	0.00	0.02	0.01
N	2,586	1,663	4,249

* *p*<0.05; ** *p*<0.01; *** *p*<0.001

Table 3.9: Linear regression model estimating the effect of one or two neighbours' defection in the PD on change in contributions of participants who contributed *less than* their neighbours previously in the PGG. Standard errors clustered on session.

	Control	Treatment	Both
1 neighbour withheld PD cooperation	-0.038 (0.365)	0.069 (0.411)	-0.038 (0.353)
2 neighbours withheld PD cooperation	0.129 (0.500)	1.981 (0.393)**	0.129 (0.483)
1=Treatment			-0.430 (0.426)
1=Treatment X 1 neighbour withheld			0.107 (0.531)
1=Treatment X 2 neighbours withheld			1.852 (0.614)**
Constant	1.869 (0.261)***	1.439 (0.355)**	1.869 (0.252)***
R^2	0.00	0.02	0.01
N	2,586	1,663	4,249

* *p*<0.05; ** *p*<0.01; *** *p*<0.001

Table 3.10: Linear regression model estimating the effect of both neighbours' cooperation in the PD on change in contributions of participants who contributed the *same or more than* their neighbours previously in the PGG. Standard errors clustered on session.

	Control	Treatment	Both
# Neighbours cooperating in PD	0.253	0.828	0.253
	(0.171)	(0.132)***	(0.166)
1=Treatment			0.319
			(0.290)
1=Treatment X neighbours cooperating			0.575
			(0.209)*
Constant	-2.523	-2.205	-2.523
	(0.217)***	(0.208)***	(0.209)***
R^2	0.00	0.02	0.02
N	3,061	3,596	6,657

* p<0.05; ** p<0.01; *** p<0.001

3.2.2.5 Scalability

3.2.2.5.1 Random variation of group size

Finally, we present evidence that our "local-to-global" reciprocity is scalable across different sized groups. We take advantage of random variation across sessions in the number of participants in the PGG to illustrate this. One might worry that as groups become larger, local interactions between just two neighbours might be ineffective at stabilising contributions in the global PGG.

However, we find no evidence for this: contributions in the final round of the game do not decline as groups become larger in the treatment condition (coeff = -0.015, p = 0.782, Table 3.11 col. 2; all regressions in this section take the group as the unit of observation, with one data point per group). In fact, a threefold increase in the size of the group has no discernible impact on PGG contributions in the treatment.

In contrast, final round contributions in the control do seem to decrease as groups grow larger, albeit only at a marginal level of statistical significance (coeff = -0.110, *p* = 0.069, Table 3.11 col. 1). When data from both the control and treatment conditions are taken together, we correspondingly observe a positive interaction between a treatment dummy and group size (coeff = 0.095, *p* = 0.204, Table 3.11 col. 3), suggesting that our intervention if anything becomes more effective relative to the control as groups becomes larger, although this interaction does not achieve statistical significance (perhaps not surprisingly given that we have only 8 independent observations per condition and thus the statistical test has little power). Results when considering average PGG contributions over all rounds are qualitatively similar (Table 3.12).

Table 3.11: Linear regression model estimating the effect of group size on PGG contributions in the final round of the game. Each session corresponds to one observation, and robust standard errors are used.

	Control	Treatment	Both
Group size	-0.110	-0.015	-0.110
	(0.050)	(0.051)	(0.050)*
1=Treatment			3.521
			(3.077)
1=Treatment X group size			0.095
			(0.071)
Constant	11.125	14.646	11.125
	(2.469)**	(1.836)***	(2.469)***
R^2	0.36	0.01	0.85
N	8	8	16

* *p*<0.05; ** *p*<0.01; *** *p*<0.001

Table 3.12: Linear regression model estimating the effect of group size on PGG contributions averaged over all rounds of the game. Each session corresponds to one observation, and robust standard errors are used.

	Control	Treatment	Both
Group size	-0.052	-0.011	-0.052
	(0.035)	(0.029)	(0.035)
1=Treatment			-0.318
			(2.003)
1=Treatment X group size			0.041
			(0.046)
Constant	15.030	14.712	15.030
	(1.672)***	(1.103)***	(1.672)***
R^2	0.23	0.02	0.37
N	8	8	16

* p<0.05; ** p<0.01; *** p<0.001

3.2.2.5.2 Large-scale PGG with 1,000 players

To further address the question of scalability, we ran an additional experiment with 1,000 participants playing one large PGG. In the treatment condition, participants were able to see the contributions of the player with whom they also played a repeated PD after each round of the PGG. Conversely, in the control condition, participants were not able to see the contributions of their PD partner. However, each player in the control condition saw the contributions of another player who was simultaneously playing the same game with someone else (see Section 3.1 for experimental details).

This design required four participants playing the game simultaneously in two pairs in the control condition, and the decisions between those pairs are not independent. Thus, to account for this interdependence, we cluster standard errors at this "double pairs" level (i.e., four players in two pairs playing the game simultaneously). To keep decision decision times and dropout rates constant between control and treatment, we also required that two pairs of participants played the game simultaneously in the treatment group and thus we also cluster on double pairs.

As before, we predicted that contributions would be lower in control than treatment, and this difference would emerge over time. This is indeed what we found: overall levels of contribution were lower in control than treatment (coeff = -1.456, p = 0.005, Table 3.13 col. 1). Over time, participants in the control condition decreased their contributions (coeff = -0.136, $p <$ 0.001, Table 3.13 col. 2), while contributions remained stable in treatment (coeff = -0.027, p = 0.429, Table 3.13 col. 3). This difference was significant when we combined the data from both conditions (interaction between number of rounds and control dummy, coeff = -0.109, p = 0.017, Table 3.13 col. 4). Furthermore, we found qualitatively similar results when we include dropout groups in our analysis (Table 3.13 col. 5).

Table 3.13: Linear regression model estimating the effect of experimental condition and round on contributions in the large public goods game. Standard errors clustered on "double pairs" (groups of four simultaneous players).

	Combined	Control	Treatment	Interaction	Interaction
1=Control	-1.456 (0.514)**			-0.855 (0.504)	-0.886 (0.431)*
Round		-0.136 (0.030)***	-0.027 (0.034)	-0.027 (0.034)	-0.044 (0.036)
1=Control X Round				-0.109 (0.046)*	-0.098 (0.049)*
Constant	11.912 (0.362)***	11.206 (0.361)***	12.061 (0.354)***	12.061 (0.353)***	12.194 (0.307)***
Dropouts included	No	No	No	No	Yes
R^2	0.01	0.00	0.00	0.01	0.01
N	10,000	5,000	5,000	10,000	11,620

* p<0.05; ** p<0.01; *** p<0.001

Chapter 4.

Cooperating with the future

4.1 Main text

Providing for future generations is central to the survival of genes, families, organizations, nations and the global ecosystem (Hardin 1968; Ostrom 1990; Levin 2007; Milinski and Semmann 2006; Wade-Benzoni and Tost 2009). Yet providing for the future poses a challenge, as it requires making sacrifices today. Institutions can play an important role in promoting such cooperative behaviour among large groups of people. Traditionally, institutional designers have assumed that people are rational and purely self-interested, and proposed incentives that induce selfish people to cooperate (Coase 1960; Mueller 1979; Williamson 1985).

In recent years, however, a large body of evidence has demonstrated that many people are not purely selfish (Wade-Benzoni and Tost 2009; Forsythe et al. 1994; Camerer 2003; Charness and Rabin 2002; Fosgaard, Hansen, and Wengström 2011; Amir et al. 2012; Rand, Greene, and Nowak 2012). Here we consider the implications of these 'social preferences' for designing institutions that promote sustainability and intergenerational cooperation. We demonstrate that democracy can be a powerful institution for harnessing social preferences: while selfish people would vote for over-exploitation of resources, voting allows a prosocial majority to override a selfish minority. (See Supplementary Information, SI, Section 1 for further motivating discussion.)

To do so, we introduce a laboratory model of cooperating with the future – the Intergenerational Goods Game (IGG) – that builds on previous work using Public Goods Games (Milinski et al. 2001; Fehr and Gächter 2002; Rand et al. 2009), Common Pool Resource games

(Ostrom, Walker, and Gardner 1992; Walker, Gardner, and Herr 2000) and Threshold games (Milinski and Semmann 2006; Jacquet et al. 2013; Cadsby and Maynes 1998). In these other games, selfishness creates social efficiency losses for the other members of one's group. In contrast, the IGG is designed such that selfishness instead negatively impacts *subsequent* groups.

In our IGG experiments, individuals form groups of five, which we refer to as generations. The first generation is endowed with a common pool of 100 units and each individual can extract between 0 and 20 units from the pool. If the total percentage of units extracted from the pool is at or below a commonly known extraction threshold, *T*, the pool will renew to 100 units for the next generation. If, however, the percentage extracted is above *T*, the pool is exhausted and all future generations receive no payoff (Figure 4.1). After each generation, another generation occurs with probability δ, and with probability 1-δ the game ends: the discount factor δ models the extent to which the current generation values the next generation. (See Section 4.2.2 for further experimental details.)

In the game theoretic tradition, the IGG framework is a great simplification relative to real-world intergenerational cooperation. For discussion of important aspects of intergenerational transfer which the IGG does yet not incorporate, as well as relation of our work to previous results on intergenerational transfer, see Section 4.2.3.

To explore behaviour in the IGG, we begin with an 'unregulated' treatment: each group member individually chooses how many units to extract from the pool. We initialize 20 unregulated IGGs, and pass each game's pool across a series of generations with a discount factor of δ=0.8 (leading to an expected game length of five generations). For the pool to be replenished, each generation must extract 50 units or less (*T*=50%). Thus the socially efficient extraction (or 'fair share') is 10 units per individual on average. We focus on symmetric strategies and refer to individuals who extract 10 or fewer units as cooperators, and those who extract more than 10 units as defectors.

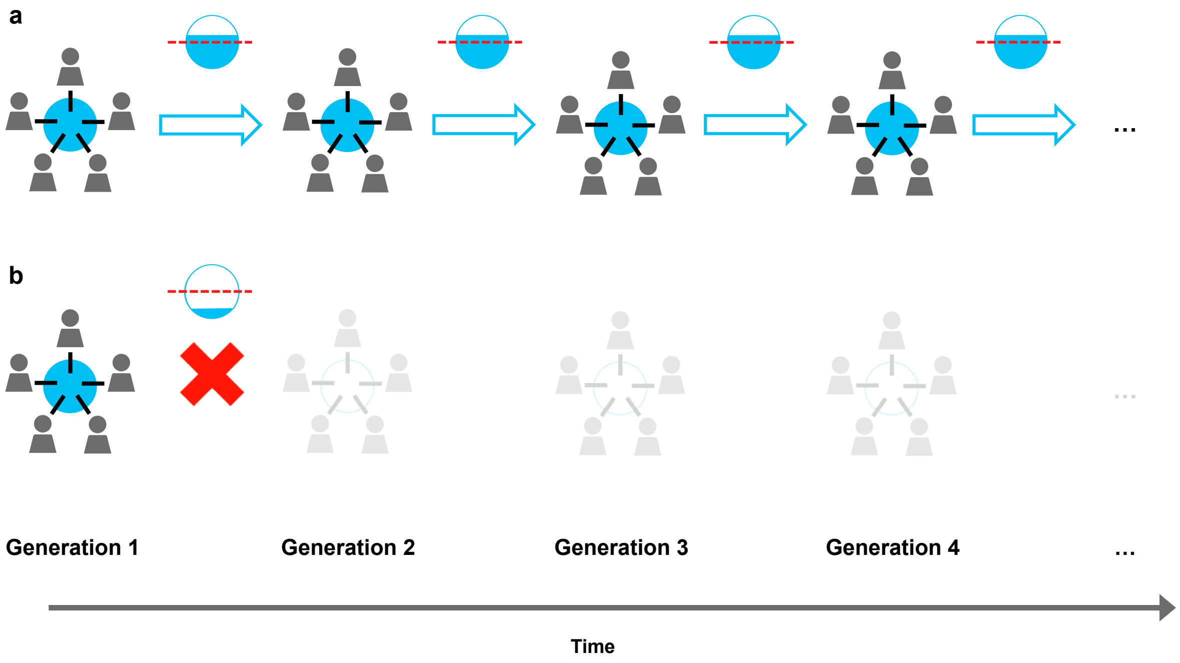

***Figure 4.1. An illustration of the intergenerational game (IGG):** In each generation, a group of 5 people makes a decision (individually or according to an institutional rule) about their level of extraction from a common resource. **a** If Generation 1's extractions do not violate the commonly known threshold, the resource refills and the same dilemma is presented to Generation 2. After each generation, another generation occurs with probability δ. **b** If at any point the threshold requirement is not met, the resource does not renew and future generations receive no payoff. Maximal social welfare is achieved if no generation ever violates the threshold requirement by extracting too much from the common resource.*

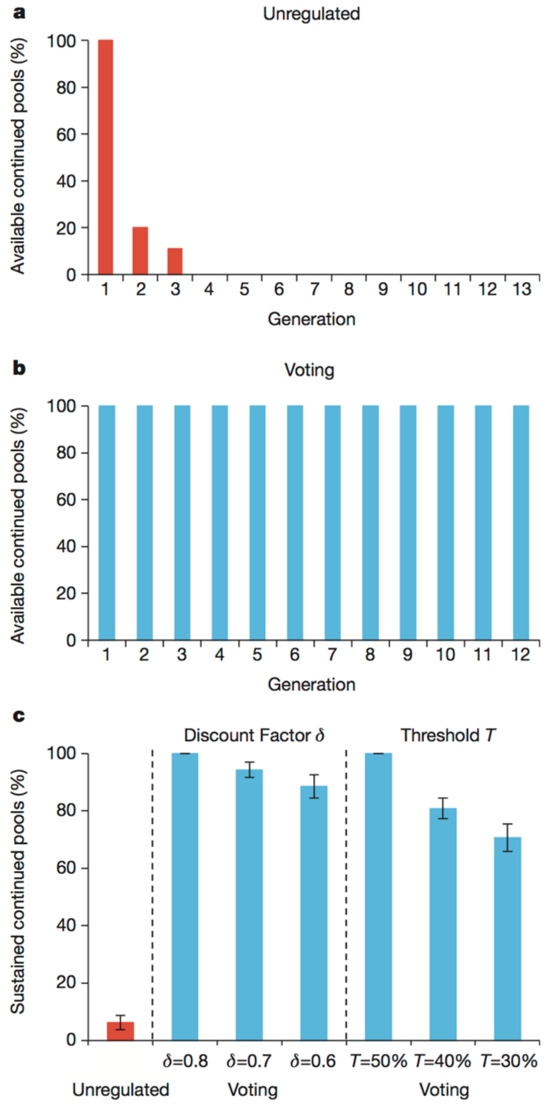

***Figure 4.2. Solving the (intergenerational) tragedy of the commons through an institutional design. a** When decisions are made at the individual level, the availability of the common pools drastically decreases over time; N=480. **b** The introduction of a democratic voting institution dramatically improves sustainability; N=370. **c** Decreasing the discount factor from δ=0.8 to δ=0.7 (N=355) or δ=0.6 (N=305) while holding T=50%, or the extraction threshold from T=50% to T=40% (N=600) or T=30% (N=460) while holding δ=0.8, increases the temptation to defect. Nonetheless, much less is extracted under median voting compared to the unregulated baseline. Errors bars indicate standard errors of the mean.*

We find that a large majority of individuals cooperate (68%), in line with previous studies using non-student populations (Fosgaard, Hansen, and Wengström 2011; Amir et al. 2012; Rand, Greene, and Nowak 2012). Despite their good intentions, however, only 4 of the 18 games continuing on to a second generation have their pools sustained. These losses in sustainability compound quickly over time: in generation 3, the number of refilled pools is down to two, and not a single refilled pool is available to the 4th generation (Figure 4.2a). Notably, in most groups, only a minority of defectors is responsible for the exhaustion of the resource.

To address this sustainability failure, we introduce an institution that is central to the Western world: democracy. Each group member votes for their generation's extraction level, and the median vote is extracted by all players. Well studied by economists and political scientists (Holcombe 1989; Walker, Gardner, and Herr 2000; Ertan, Page, and Putterman 2009; Putterman, Tyran, and Kamei 2011; Kamei, Putterman, and Tyran 2014; Bernard et al. 2013), this 'median voting' rule guarantees socially optimal outcomes in a standard Public Goods Game, even with perfectly self-interested actors: the payoff-maximising vote is full cooperation (Kamei, Putterman, and Tyran 2014; Bernard et al. 2013). In the IGG, however, this is not true: because the current group does not reap the benefits of cooperation, selfish players would vote to deplete the resource fully. From a traditional 'public choice' perspective based on rational self-interest, therefore, median voting is not attractive for promoting sustainability. If, however, enough players have social preferences, voting may be able to support sustainability in the IGG by allowing prosocial players to reign in selfish players. Thus a 'behavioural public choice theorem' (Ertan, Page, and Putterman 2009; Putterman, Tyran, and Kamei 2011; Kamei, Putterman, and Tyran 2014) might favour median voting; see Section 4.2.1 for further discussion.

To explore the effects of median voting, we initialize another 20 IGGs using δ=0.8 and T=50%, and applied the voting rule. We find a dramatic increase in sustainability (Figure 4.2b): all 20 common pools are sustained across all generations (unregulated vs. voting: linear probability model (LPM) predicting pool sustainability at the generation level, p<0.001; see Section 4.2.4 for statistical details).

Next we ask how robust the voting mechanism is to variation in the discount factor, δ, and the extraction threshold, T. In the experiments described above, there was an 80% chance that a future generation would exist (δ=0.8) and individuals had to sacrifice half of their possible payoff to extract a 'fair share' (T=50%). We now examine the effectiveness of voting in two treatments using lower δ values (δ=0.7 and δ=0.6, creating fewer future generations), and two other treatments using lower T values (T=40% and T=30%, leading to a higher cost of cooperation). Each treatment again started with 20 pools.

We find that voting remains largely effective in promoting sustainability under these more adverse conditions (Figure 4.2c). Although sustainability does vary significantly with δ (LPM, p=0.037) and T (LPM, p<0.001), the size of these effects is relatively small: decreasing δ or T by 0.1 decreases the probability of a pool being sustained by 4.6% or 14.6%, respectively. Moreover, under all conditions tested, voting leads to much higher levels of sustainability than the original unregulated IGG (LPM, p<0.001 for all comparisons).

The success of voting is driven by two factors. First, the decision-making power differs in the voting and unregulated institutions (Figure 4.3a). In the voting institution, a majority of three cooperators who propose 10 unit extractions can overrule two defectors who propose 20 units. In contrast, if decisions are made at the individual level, a single defector can tip the balance of a group. In other words, voting allows a majority of cooperators to restrain a minority of defectors.

The second reason for the success of voting pertains to the psychology of social preferences. Median voting addresses the fears of players who care about future generations but worry that others (now or later) will exhaust the pool (i.e., conditional cooperators (Fischbacher, Gächter, and Fehr 2001)): since the outcome of the vote is applied to all players, everyone within a generation receives the same payoff and no one risks being the 'sucker'. This, in turn, further increases the probability that a cooperative majority is formed and the pool is sustained, both in the current generation and in the future. Figure 4.3b is consistent with this assessment: the fraction of cooperators is 20% larger under voting than unregulated (LPM coef=0.201, p<0.001).

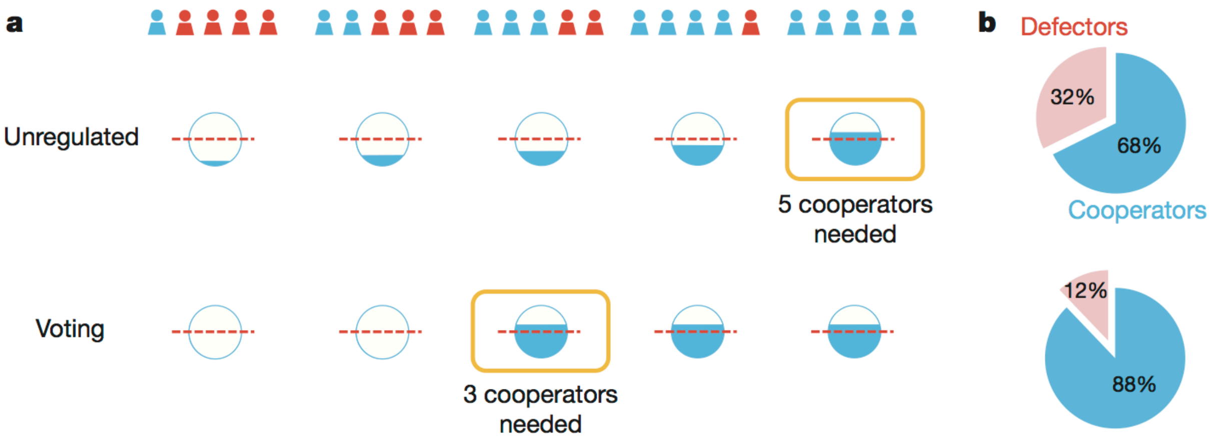

Figure 4.3. The voting institution is robust to extreme decision-makers and thereby increases cooperative behaviour. a *The pivotal decision-maker in the voting institution is different from the unregulated institution. For instance, assume that T=50%, and that a cooperator and a defector always extract 10 and 20 units, respectively. The unregulated institution is vulnerable to extreme decision-makers, whereas the voting institution is robust to a minority of defectors. This, in turn, bolsters the decision of those who are predisposed towards cooperation but fear to be exploited (e.g., future-oriented 'conditional cooperators').* ***b*** *This leads to an increase of cooperators in the voting institution (N=370) over the unregulated institution (N=480).*

Both of these factors predict that voting is only successful if everyone is bound by the outcome: a partial implementation (Bernard et al. 2013) provides an opportunity both for defectors to derail sustainability, and for potential cooperators to switch to defection out of fear that others will over-exploit.

We test this prediction by introducing a 'partial voting' treatment (another 20 pools, again using δ=0.8 and T=50%). Three of the five people in each generation are bound by the decision of a median vote among themselves. The other two people are not informed of the vote's outcome, and decide freely how much to extract. The sum of all five extractions is then compared to the extraction threshold T.

As predicted, the partial voting institution is significantly less successful than the full voting institution (Figure 4.4a, LPM p<0.001). This point is driven home by bootstrapping simulations: of 10,000 pools created by randomly sampling participant decisions each generation, only 1.5% of available pools are sustained after 15 generations under partial voting, compared to 84% under full voting; see Figure 4.5 and Section 4.2.5 for details. We conclude that, for voting to effectively manage sustainability, it must be binding for all decision-makers.

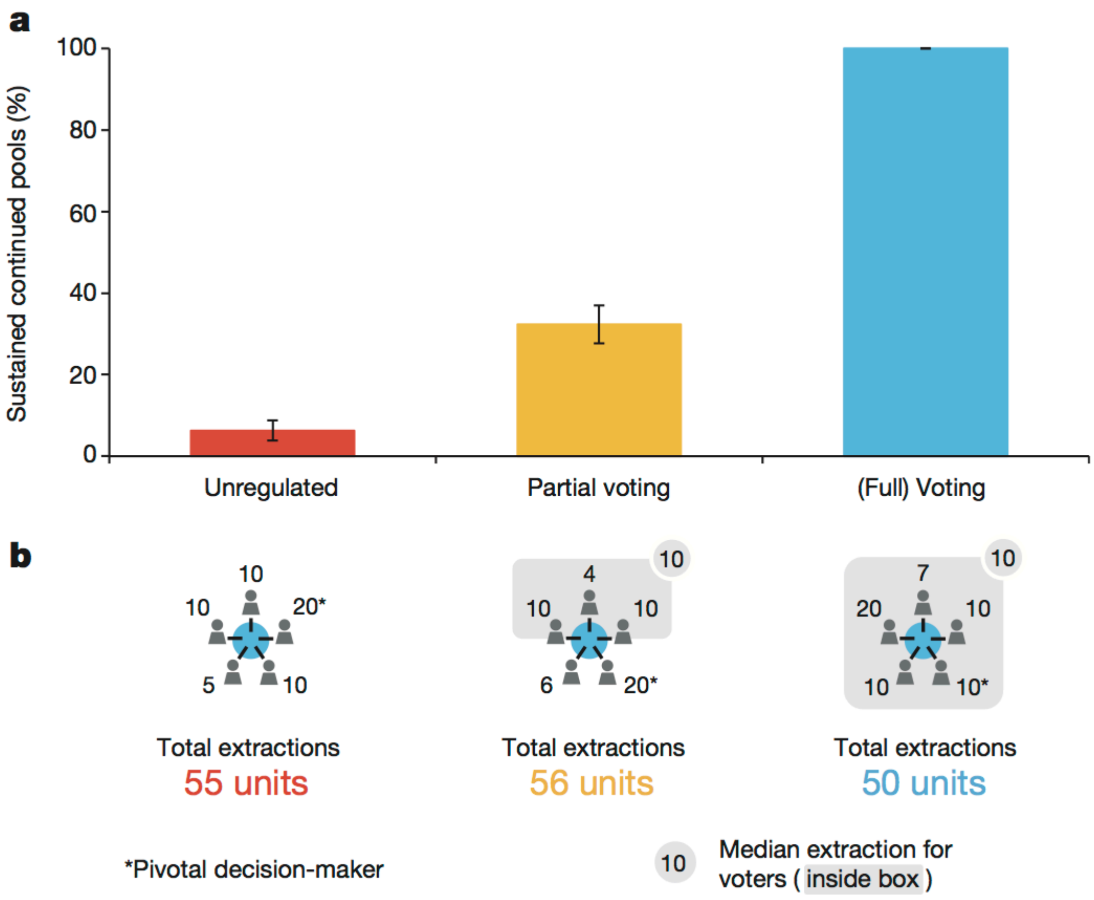

***Figure 4.4. We confirm the hypothesis that voting must be binding for all players to achieve high levels of sustainability. a** In a partially implemented voting institution (N=495), three of the individuals are bound to a vote while the other two can extract at will. A partially implemented voting institution is not robust to a minority of defectors and also cannot reassure conditional cooperators. Thus, partial voting fails to lead to sustainable outcomes.* ***b** Three real sets of decisions from our data demonstrate a consequence of the pivotal extractor outside the voting group.*

In this paper, we have introduced a new laboratory model for cooperation across generations, the Intergenerational Goods Game (IGG). We have shown that in the absence of regulation, a minority of selfish players consistently deplete available resources. By implementing median voting, however, this negative outcome can be prevented – but only if all players are bound by the outcome of the vote. Votes that are only partially binding, such as the international Kyoto protocol, have little power.

More generally, our results emphasize the importance of institutional designers moving away from the assumption of universal self-interest. We extend the 'behavioural public choice theorem' (Ertan, Page, and Putterman 2009; Putterman, Tyran, and Kamei 2011; Kamei, Putterman, and Tyran 2014) by demonstrating how voting can allow a majority of prosocial individuals to override a purely selfish minority, leading to costly group-level cooperation with future generations. Real-world data are consistent with this suggestion: countries that are more democratic also have more sustainable energy policies (combining data for 128 countries from the Economist's Democracy Index and World Energy's Energy Sustainability Index, $p<0.001$, $R^2=0.36$; robust to controlling for GDP, Gini index, population size, literacy rate, unemployment rate, life expectancy, and level of corruption; see Figure 4.6 and Section 4.2.6 for details). Policy makers can do much to promote the public good by using a behavioural approach that is informed by a more accurate understanding of human psychology (Kamei, Putterman, and Tyran 2014; Oullier 2013; Benkler 2011; Haynes et al. 2012). Many citizens are ready to sacrifice for the greater good. We just need institutions that help them do so.

4.2 Supporting figures and data

4.2.1. Theoretical motivation

In this paper, we ask how different institutional rules lead individuals to conserve resources, leaving enough to provide for the next generation. In particular, we are interested in institutions that create sustainable outcomes by harnessing social preferences, and thus may be overlooked when relying on assumptions of rational self-interest.

We focus on the institution of median voting (Holcombe 1989; Deacon and Shapiro 1975; Walker, Gardner, and Herr 2000; Putterman, Tyran, and Kamei 2011). Among selfish players, median voting can promote *intra*-generational cooperation (i.e. cooperation in traditional public goods and common pool resource games) (Walker, Gardner, and Herr 2000; Bernard et al. 2013). The essential structure of *intra*-generational cooperation is that a group of cooperators earns more than a group of defectors, but that the highest payoff comes from unilaterally defecting in a group where everyone else cooperates. Because median voting binds all players to the same action, unilateral defection is impossible. Therefore the highest payoff is earned by being in a group where everyone cooperates, and selfish players will vote for cooperation. This makes median voting an attractive institution for promoting cooperation under assumptions of rational self-interest in an intra-generational social dilemma.

However, this is not true in the context of *inter*generational cooperation. In our intergenerational goods game (IGG), all benefits created by the current generation's cooperation are reaped by subsequent generations. Therefore it is no longer true that a group of cooperators earns more than a group of defectors. Instead, it is the case that a series of cooperative groups, who sustain a pool over multiple generations, earn more in *total* than a series where one defecting groups exhausts the pool early on. But an *individual*'s payoff is unaffected by the choices of the other members of her own generation. Whereas in the traditional PGG, one's own payoff increases when others in the same group cooperate, in the IGG one's own payoff increases when members of the previous generation have cooperated.

Therefore a player in generation *i* maximizes her payoff in the IGG by extracting the maximum amount, and is indifferent (monetarily) to the extraction amounts of other members of generation *i*. Because of this, selfish players will vote to extract the maximum amount in the IGG, unlike in the PGG. This would lead traditional theories of public choice, based on rational self-interest, to conclude that median voting is not a good solution for promoting intergenerational transfer.

However, the picture changes dramatically once social preferences are taken into account. A large body of literature suggests that a majority of people in many contexts are not purely self-interested, but instead care to some extent about the well-being of others (Camerer 2003). People with these kinds of prosocial preferences may be willing to pay a cost to benefit members of future generations. However, they may also have 'conditional cooperation' preferences (Fischbacher, Gächter, and Fehr 2001), which is to say that they prefer cooperating as long as others (both in their own generation and in future generations) cooperate as well.

Consider how a strong conditional cooperator who cares about future generations would play our IGG (in groups of five). In the unregulated condition, she would cooperate if she expected all four others to cooperate (and all members of future generations to cooperate), and would defect otherwise. Under median voting, however, she would vote for cooperation as long as she expects at least two others in her generation (and 3 others in future generations) to vote for cooperation, because only three cooperative votes are needed to make the median vote cooperative. Thus, in a population of players of which some are future-oriented conditional cooperators, median voting can substantially increase the fraction of people *choosing* to cooperate.

Furthermore, median voting decreases the number of cooperators *needed* for sustainability to be achieved. Following similar logic as above, five cooperative choices are needed in the unregulated cases, whereas only three are needed under median voting. Therefore, median voting also makes it easier not to over-exploit the pool.

Critically, however, the predicted success of median voting hinges on a large fraction of the population having social preferences. If all players were purely self-interested, sustainability would never be achieved in either the unregulated or the median voting conditions.

Our experiments are therefore designed to differentiate between the pessimistic prediction of classical public choice theory based on rational self-interest, and the optimistic prediction of a 'behavioural public choice theorem' rooted in social preferences (Kamei, Putterman, and Tyran 2014).

4.2.2. Methods

4.2.2.1 Data Collection on Amazon Mechanical Turk

For all of our experiments, we recruited U.S. residents to participate using the online labour market Amazon Mechanical Turk (AMT). Our experiment was approved by Harvard University Committee on the Use of Human Subjects in Research, and informed consent was obtained from all subjects.

To preserve random assignment, each generation for all conditions was run at the same time, and subjects within each generation were randomly assigned to one of our seven experimental condition. Each experimental condition is described in more detail below.

AMT is an online market place in which employers can pay users for completing short tasks (generally about 10 minutes) – usually referred to as Human Intelligence Tasks (HITs) – for a relatively small pay (generally less than a $1). Workers who have been recruited on AMT receive a baseline payment and can also be paid a bonus depending on their performance in the task. This setup lends itself well to adopt incentivised economic experiments: the baseline payment acts as the 'show up' fee and the bonus payment may derive from the workers' behaviour in the economic game and/or other tasks throughout the experiment.

A major advantage for using AMT is that the sample of recruited subjects has been shown to be more diverse and more nationally representative than the typical college student sample at major research universities, at which many economic games are run (Buhrmester, Kwang, and Gosling 2011; Amir et al. 2012; Horton, Rand, and Zeckhauser 2011).

There may, of course, exist potential issues on AMT that would not occur in a traditional laboratory setting. For instance, running an experiment online involves giving up some control over subjects, since they cannot be monitored, as is usually the case in laboratories. That is, it cannot be ruled out that more than a single person is taking part in the experiment or that one person is participating more than once in the experiment (although AMT has put extensive measures into place to avoid this from happening; in addition, we have also implemented ways to carefully screen out any possible re-takers). Finally, the participating subject sample, albeit more diverse and representative than the average college students sample, is biased towards those who participate in online labour markets in the first place. To address these possible concerns, numerous studies have been carried out to validate results collected using AMT. Of particular relevance to the present study, very similar levels of prosociality have been found on AMT and in the lab (using an order of magnitude higher stakes) in a one-shot Prisoner's Dilemma (Horton, Rand, and Zeckhauser 2011), Public Goods Game (Amir et al. 2012), Trust Game (Amir et al. 2012) and Ultimatum Game (Amir et al. 2012).

4.2.2.2 Basic flow of the experiments

All participants earned a $0.50 showup fee and had the opportunity to earn up to an additional $1 in bonus payments depending on the outcome of the IGG. Participants took part in the experiment through an online survey provided by Qualtrics. After participants had read the experiment instructions (see Section 4.2.6 below), they had to pass a comprehension quiz about the rules of the game in order to partake in the actual experiment. Those who didn't pass the quiz received only the baseline payment of $0.50, and are excluded from our analyses (in accordance with common practice on AMT (Horton, Rand, and Zeckhauser 2011)).

The details of the decision-making stage depended on experimental condition, and the state of the common pool: if the threshold requirement had not been violated by a previous group and the pool had thus refilled to 100 units, participants made their choice about their extraction or vote (depending on the experimental condition they were in; explained in more detailed below). If the extraction threshold *had* been exceeded previously, the common pool was empty; in this case, the participants were informed about this fact and made no decision nor received any bonus payment beyond the baseline payment (i.e., the show-up fee).

We were concerned that the instructions in the Partial Voting condition were more complicated than in the other conditions, and therefore that substantially more subjects might fail the initial comprehension check in Partial Voting (and thus be excluded). This could be potentially problematic because it would mean that the people who got to participate in the Partial Voting condition would be 'smarter' on average than participants in the other conditions, and this could bias our results if ability to pass more complicated comprehension questions was correlated with behaviour in the IGG.

We sought to mitigate this problem as follows. *After* the decision-making stage (so as to not influence their decisions), participants in all conditions other than Partial Voting were presented with the Partial Voting instructions and the corresponding comprehension quiz. Participants had to pass both comprehension checks correctly to receive their bonus, and we only include subjects who passed both quizzes in our analyses. (To keep the experiment approximately equal length across conditions, we also required participants in the Partial Voting condition to read the instructions for the Unregulated condition and answer the associated comprehension questions.)

As expected, substantially fewer subjects passed the first set of comprehension questions in the Partial Voting condition (52%) compared to the other simpler conditions (67%). However, our mitigation strategy was largely successful: the fraction of subjects passing *both* sets of comprehension questions was much closer, with 49% passing Partial and 54% passing in the other conditions. Given the considerable magnitude of our treatment effects reported below, we think it is unlikely that this 5% difference in comprehension rates had a substantially effect on our results.

The experiments were approved by the Harvard University Committee on the Use of Human Subjects in Research.

4.2.2.3 General experimental design

In total our experiment had seven experimental conditions: unregulated, baseline voting, voting δ=0.7, voting δ=0.6, voting T=40%, voting T=30%, partial voting. Before describing the details of each condition, we describe the basic structure which is common to all conditions.

In each condition, 20 resource pools were initiated with 100 units in generation 1. After every generation, there was a probability *δ* that another generation would be recruited (i.e. that the game would be 'continued'). In most conditions, *δ = 0.8*, such that the expected number of generations per game was 5. We chose a sample size of 20 games per condition at the outset of the experiment, and did not collect any additional data once all 20 games had been run.

Generations were recruited sequentially, with each generation being informed of the outcome of the previous generation as described below. The list of resulting lengths (i.e. # of generations) for each game in each treatment was:

Unregulated	8	2	6	3	4	12	2	2	13	3	6	1	5	4	6	1	5	4	5	4
Voting baseline	5	2	1	8	5	5	3	5	2	1	12	1	4	5	2	3	1	2	5	2
Voting *δ*=0.7	4	1	1	2	5	1	3	2	6	2	2	2	5	3	13	2	9	1	1	6
Voting *δ*=0.6	6	1	2	1	1	2	5	8	1	2	2	1	1	2	8	6	3	4	4	1
Voting *T*=40%	4	1	12	4	1	12	4	2	6	14	13	5	1	7	12	4	1	5	7	5
Voting *T*=30%	5	4	1	6	6	6	3	5	6	4	1	1	3	5	6	13	4	10	2	1
Partial Voting	13	1	1	3	2	14	5	1	11	2	3	5	9	1	5	7	1	4	9	2

Within each generation of a game, a group of five participants chose how many units to extract from the pool (out of a total of 100 units). (The mechanism by which this choice was made varied across conditions, as described below). If the fraction extracted within a given group did not exceed the extraction threshold of *T*, that group's pool would be 'sustained': the next generation would receive a pool refilled to 100 units and have a chance to make their own set of extraction decisions (provided that the game was continued based on the continuation probability δ such that there was indeed another generation).

If, on the other hand, the fraction extracted exceeded T, the pool was exhausted. All future generations were informed that previous more than T units had been extracted from the pool, and as a result they (the current generation) would not have the opportunity to play the IGG or receive any bonus payment.

Participants in Generation 1 were informed that they were the first generation. Participants in subsequent generations were informed that the previous generation had either sustained or not sustained the pool. They were not informed, however, of the specific generation number (other than showing that they were not the first) they were because the total number of continued generations varied across pools and conditions due to the random continuation device, and we did not want to introduce this as a source of bias.

Note that in the IGG, a series of generations where each generation acts sustainably (i.e., extracts T units) has an expected total payoff of $T/(1-\delta)$. A series of generations where everyone extracts the maximum (and therefore the pool is exhausted after the first generation), has an expected total payoff of 100 (the contents of the pool in the first generation). Thus acting sustainably is socially efficient as long as $T/(1-\delta) > 100$.

4.2.2.4 Details of each condition

Our experimental conditions differed in two ways: the manner in which the number of extracted units was determined (i.e., the institution: unregulated, voting, or partial voting) and the specific values of δ and T.

First we describe the three different institutions.

- Under the unregulated institution, each of the five group members independently selected an extraction amount between 0 and 20 units.
- Under the voting institution, each of the five group members proposed an extraction amount between 0 and 20 units. The median proposal amount was then extracted for each group member.
- Under the partial voting institution, (i) three of the five group members proposed an extraction amount between 0 and 20 units and the median proposal amount was then extracted for each of the three; while (ii) the other two group members independently selected an extraction amount between 0 and 20 units.

The values of δ and T used in condition are given below, as well as the average game length and resulting number of subjects recruited (note that the criterion for sustainability to be socially efficient, $T/(1-\delta) > 100$, is met in all cases):

Condition	δ	T	***Mean # Generations***	N
Unregulated	0.8	50%	4.8	480
Voting	0.8	50%	3.7	370
Voting $\delta = 0.7$	0.7	50%	3.6	355
Voting $\delta = 0.6$	0.6	50%	3.1	305
Voting $T = 40\%$	0.8	40%	6.0	600
Voting $T = 30\%$	0.8	30%	4.6	460
Partial voting	0.8	50%	5.0	495

4.2.2.5 Statistical analysis

To analysis the data, we use linear probability models to estimate the effect of institution on pool sustainability. Groups in which the game was continued and the pool was sustained are coded as 1 (i.e., groups which received a full pool and did not extract more than *T* units). Groups in which the game was continued but the pool was exhausted are coded as 0 (i.e., groups which received a full pool and extracted more than *T* units, or groups which received an exhausted pool). Once a game was discontinued (i.e., after the random number drawn for continuation was greater than δ), no more groups were recruited and so no subsequent generations for that pool appear in the regression.

Thus we compare the fraction of continued games that have sustained pools, with one observation per group of five participants. Because of the randomness of the continuation device, the number of groups is not identical across conditions. To partially address this issue, as well as to account for the fact that outcomes of groups which receive exhausted pools are not independent of outcomes of earlier groups in that game, we cluster standard errors in our regressions at the level of the game. Furthermore, in Section 4.2.4, we compliment this regression analysis with an analysis using a large number of simulated game lengths and random permutations of extraction decisions/proposals.

4.2.3. Future directions for the IGG and relations to previous work

In the game theoretic tradition, the IGG framework is a great simplification relative to real-world intergenerational cooperation. We feel that this simplification captures key elements of the intergenerational challenge facing our world: the game is non-zero sum, with cooperation today creating greater benefits for the future; the consequences of consumption are non-linear, such that some amount of consumption can occur in the present without imposing costs on the future; and the cooperative challenge involves group-level decisions rather than just individual-to-individual transfers.

There are, however, many aspects of intergenerational cooperation which the IGG does yet not incorporate. Although we include a probabilistic continuation rule for future discounting, there may be important elements of the psychology of discounting which this approach does not capture. Additionally, future generations are our relatives, will likely be richer than us, and may have access to technological innovations that could mitigate our current environmental concerns.

Future work extending our IGG framework to examine these issues, as well as exploring intergenerational cooperation among larger groups, overlapping generations (Van der Heijden and Nelissen 1998), and groups with the possibility of borrowing against the future (i.e. running up national debt) will help to advance our understanding of real-world intergenerational cooperation. So too will considering spatial effects, where over-consumption in one area has little consequence for individuals living far away (Janssen et al. 2010).

Our IGG experiments add to a nascent literature on cooperation across generations. Previous work has demonstrated that coordination, communication and social reputation help meet targets in groups to avoid collective loss (Milinski and Sommerfeld 2008; Milinski and Semmann 2006; Tavoni, Dannenberg, and Kallis 2011; Milinski, Röhl, and Marotzke 2011). Our voting intervention is also a type of coordination mechanism. It helps coordinate people’s preferences towards their own gains and those of future generations (Jacquet et al. 2013).

Additionally, other work has emphasised that altruism depends on previous generations’ behaviour as well as the personal distance between donors and recipients of the intergenerational good (Wade-Benzoni 2002; Wade-Benzoni and Tost 2009). We thus expect that the results in our

IGG would be further magnified if the longevity of a common pool was made salient to later generations or if personal relationships existed between individual members of generations, as in the case of families. In contrast, we expect that factors such as uncertainty, inequality, and global sanctioning approaches would lead to lower rates of cooperation and sustainability (Tavoni, Dannenberg, and Kallis 2011; Barrett and Dannenberg 2012; Vasconcelos, Santos, and Pacheco 2013).

Our experiments also build on previous work exploring the interaction between voting institutions and social preferences. In particular, it has been shown that voting mechanisms can override anti-social behaviour where cooperators are punished, because typically only a minority hold such anti-social preferences (Putterman, Tyran, and Kamei 2011; Kamei, Putterman, and Tyran 2014; Ertan, Page, and Putterman 2009). Our results extend this 'behavioural public choice theorem'. We demonstrate how voting can allow a majority of prosocial individuals to override a purely selfish (rather than anti-social) minority, leading to costly group-level cooperation with future generations.

4.2.4. Statistical details

4.2.4.1 Unregulated vs. voting

Here we ask the basic question of how sustainability under the unregulated institution compares to sustainability under the voting institution (both using δ = 0.8 and T = 50%). We begin by considering just the first generation (Table 4.1 col. 1). We see that dramatically more pools are sustained under voting. Pooling across all generations (Table 4.1 col. 2) we see an even bigger positive effect of voting.

Table 4.1: Linear probability model estimating the effect of institution on pool sustainability. Standard errors clustered at pool level.

	1st Generation	All Generations
1=Voting	0.800	0.938
	(0.092)***	(0.025)***
Constant	0.200	0.062
	(0.092)*	(0.025)*
R^2	0.67	0.87
N	40	170

* p<0.05; ** p<0.01; *** p<0.001

In addition to this group-level outcome, we examine how the voting institution changes behaviour at the individual level. In particular, we examine the fraction of subjects behaving prosocially in each condition.

To do so, we label individuals as “cooperators” if they choose to extract 10 units or less in the unregulated condition, or vote to extract 10 units or less in the voting condition. We then use a linear probability model to estimate the effect of institution on proportion of cooperators. Both in the first generation (Table 4.2 col. 1) and over all generations (Table 4.2 col. 2), significantly more participants are cooperators in the voting condition than the unregulated treatment.

To demonstrate that this finding is not an artefact of our binary classification of subjects as Cooperators or Non-cooperators, we also estimate the effect of institution on participants’ decision (extraction amount in unregulated, proposal amount in voting). Consistent with the binary analysis, participants’ decision extraction amounts are significantly lower under voting than when unregulated, both in the first generation (Table 4.2 col. 3) and over all generations (Table 4.2 col. 4).

Table 4.2: Linear probability model estimating the effect of institution on likelihood of cooperation (col 1 and 2). Linear regression estimating the effect of institution on average decision (col 3 and 4). Standard errors clustered at the pool level.

	Cooperator?		Decision/Proposal	
	1^{st} Generation	All Generations	1^{st} Generation	All Generations
1=Voting	0.220 (0.062)***	0.201 (0.050)***	-1.980 (0.784)*	-2.290 (0.634)***
Constant	0.660 (0.054)***	0.677 (0.048)***	11.480 (0.677)***	11.485 (0.589)***
R^2	0.07	0.05	0.03	0.04
N	200	500	200	500

* p<0.05; ** p<0.01; *** p<0.001

4.2.4.2 Effects of reducing δ and T

Next we ask how reducing the discount factor δ and the extraction threshold *T* affects sustainability in the voting institution. To do so, we analyse all data from the five voting conditions (but not the ‘partial voting’ condition) jointly in a linear probability model, and estimate the probability of pools being sustained.

Examining just the first generation (Table 4.3 col. 1), we see that neither *δ* (p = 0.223) nor *T* (p = 0.441) significantly affect sustainability, although both effects are trending in the positive direction (i.e. lower *δ* and lower *T* lead to less sustainability). Examining all generations (Table 4.3 col. 2), these effects accumulate, and we do observe significant decreases in sustainability when decreasing either *δ* (p = 0.037) or *T* (p < 0.001). However, the size of these effects is not so large quantitatively: decreasing *δ* by 0.1 decreases the probability of a pool being sustained by 4.6%; and decreasing *T* by 10% decreases the probability of a pool being sustained by 14.6%.

Table 4.3: Linear probability model estimating the effect of *δ* and *T* on pool sustainability under the voting institution. Standard errors clustered at pool level.

	1st Generation	All Generations
δ	0.300 (0.245)	0.460 (0.220)*
T	0.002 (0.003)	0.015 (0.002)***
Constant	0.660 (0.254)*	-0.119 (0.235)
R^2	0.01	0.08
N	100	418

* p<0.05; ** p<0.01; *** p<0.001

Table 4.4: Linear probability model comparing pool sustainability in the unregulated condition (taken as the baseline) to the voting institutions with reduced *δ* or *T*. Standard errors clustered at pool level.

	1st Generation	All Generations
1=Voting δ=0.7	0.750 (0.105)***	0.881 (0.037)***
1=Voting δ=0.6	0.750 (0.105)***	0.823 (0.048)***
1=Voting T=40%	0.750 (0.105)***	0.644 (0.054)***
1=Voting T=30%	0.800 (0.092)***	0.746 (0.044)***
Constant	0.200 (0.092)*	0.062 (0.025)*
R^2	0.61	0.46
N	100	440

* p<0.05; ** p<0.01; *** p<0.001

Most importantly, the probability of a pool being sustained under voting in any of these reduced *δ* or *T* cases is dramatically higher than when unregulated (Fraction of pools sustained: Unregulated, 6.3%; δ=0.7, 94%; δ=0.6, 89%, T=40, 81%; T=30, 71%; all differences from unregulated $p < 0.001$, see Table 4.4). Note that this is true even though the unregulated condition has the advantage of higher δ or T, depending on the voting condition.

4.2.4.3 Partial voting

Finally, we examine the effect of a partial voting institution, under which only three of the five group members are bound by a vote. We use $\delta = 0.8$ and $T = 50\%$, and compare the fraction of pools sustained to all of our previous conditions. To do so, we use a linear probability model taking partial voting as the baseline, and estimate the proportion of pools sustained including dummies for each other condition.

We see that both in the first generation (Table 4.5 col. 1) and over all generations (Table 4.5 col. 2), sustainability is dramatically lower in the partial voting condition than in any of the voting conditions, although partial voting does still lead to somewhat more sustainability than the unregulated case.

Table 4.5: Linear probability model comparing pool sustainability in the partial voting condition (taken as the baseline) to all other conditions. Standard errors clustered at pool level.

	1^{st} Generation	All Generations
1=Unregulated	-0.500	-0.261
	(0.140)***	(0.053)***
1=Voting baseline	0.300	0.677
	(0.105)**	(0.047)***
1=Voting δ=0.7	0.300	0.485
	(0.105)**	(0.060)***
1=Voting δ=0.6	0.250	0.383
	(0.116)*	(0.067)***
1=Voting T=40%	0.250	0.620
	(0.116)*	(0.055)***
1=Voting T=30%	0.250	0.562
	(0.116)*	(0.063)***
Constant	0.700	0.323
	(0.105)***	(0.047)***
R^2	0.50	0.47
N	140	613

* $p<0.05$; ** $p<0.01$; *** $p<0.001$

4.2.5. Simulated sustainability analysis

Our analyses thus far have examined the actual outcomes that occurred in our experiment: the fraction of available pools that had been sustained in each generation, across the 20 pools initialized at the start of each condition. There are numerous sources of stochasticity that introduce noise into these comparisons. Due to the random continuation probability, some conditions lasted for more generations on average than others. The particular random matching of subjects into groups of five can affect the outcome: consider five subjects that cooperate and extract 10 units in the baseline, and five other subjects that defect and extract 20 units. A random matching that puts the five cooperators together and the five defectors together results in one sustained pool and one exhausted pool. But any other matching would result in two exhausted pools. Finally, a pool which is exhausted in an early generation which is then continued for many generations results in a very low sustainability score; whereas if the same pool had been continued for only one round, it would have had a high sustainability score.

We address all of these sources of noise by conducting a set of computer simulations using the data generated by our participants. To do so, we take advantage of the fact that after the first generation, all subjects in a given condition received the same set of information, and therefore made decisions which are effectively interchangeable. Thus in each simulation run, we randomly sample (with replacement) a series of generations of participant decisions, and calculate the fraction of those generations in which the pool was refilled.

Specifically, our procedure worked as follows, for each condition:

1. For the first generation, randomly sample (with replacement) five participants from the first generation of the current experimental condition.
2. Based on their decisions, and the rules of the experimental condition, determine whether the pool is sustained or exhausted.
3. Determine if this game is continued for another generation by comparing a random number to δ.
4. If so, randomly sample (with replacement) five participants from all generations of the current experimental condition except the first generation.
5. If the pool has previously been exhausted, mark this generation as non-sustained. If the pool has not previously been exhausted, mark this generation as sustained, and determine (based on the sampled decisions and the rules of the experimental condition) whether the pool is sustained or exhausted for the next generation.

6. Determine if this game is continued for another generation by comparing a random number to δ. If so, repeated steps 4 thru 6.

Using this procedure, we first simulated 10,000 pools out to 15 generations for the Unregulated, Voting and Partial Voting conditions. As can be seen in Figure 4.5a, the results are striking: the Voting institution is dramatically more successful at sustaining the pool than either the Partial Voting or the Unregulated conditions.

We also used this procedure to examine the consequences of changing δ and T under the Voting institution. To do so, we simulated 10,000 games out to 15 generations for the T = 40% and T = 30% conditions, and 1,000,000 games out to 15 generations for the $\delta = 0.7$ and $\delta = 0.6$ conditions, and compared the results to the Voting condition simulations above. We simulated a larger number of replicates for the lower δ conditions because the games in those conditions were dramatically shorter on average, and so many more replicated were required to get a reasonable amount of data out to 15 generations. As can be seen in Figure 4.5b, reducing δ has only a small effect, and although reducing T does undermine sustainability, the effect is much less dramatic than that of Unregulated or Partial Voting despite the higher value of T in these less-regulated conditions.

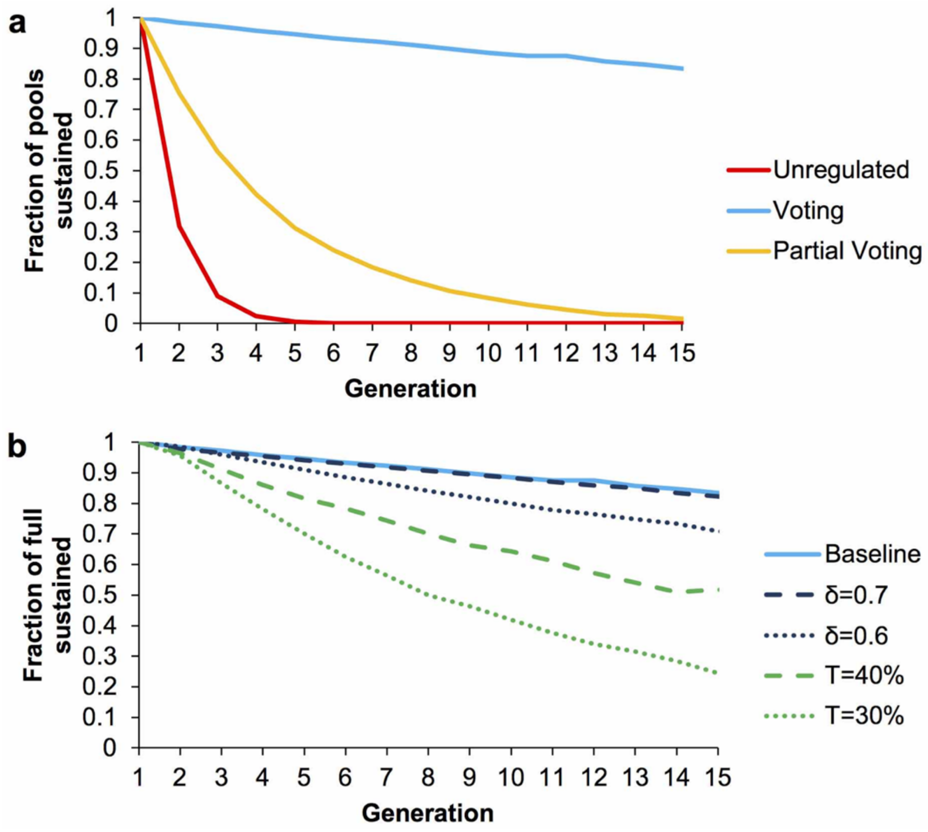

Figure 4.5. Bootstrapping simulations demonstrate the robustness of full voting, and the failure of partial voting. *We address sources of noise in the sequence of events that occurred in our experiment by conducting a set of computer simulations using the data generated by our participants. We randomly sample (with replacement) a series of generations of participant decisions, and calculate the fraction of those generations in which the pool was refilled. For each condition, we simulate 10,000 pools (or 1,000,000 pools if δ<0.8) for 15 generations.* ***a****, Simulated data for the unregulated, full voting and partial voting conditions show that the voting is by far the most successful at sustaining the pool.* ***b****, Simulated data for the T=40%, T=30%, δ=0.7 and δ=0.6 conditions shows that reducing δ has only a small effect, and although reducing T does undermine sustainability, the effect is much less dramatic than that of unregulated or partial voting despite the higher value of T in these less-regulated conditions.*

4.2.6. Positive association between democracy and sustainability

Our experiments suggest that a democratic institution can help a majority of prosocial individuals override a purely selfish minority, leading to costly group-level cooperation with future generations. Real-world data are consistent with this suggestion from our experiments: across 128 countries, more democratic institutions are associated with greater efforts to act sustainably and mitigate environmental impact.

To provide this evidence, we combine data from two independent sources: The 2012 Democracy Index (The Economist Intelligence Unit 2012) created by the Economist Intelligence Unit (EIU, part of the Economist magazine family of businesses) and the 2013 Sustainability Index created by the organization World Energy (WorldEnergy.org, n.d.). Both data files are publicly accessible on the respective websites. We examine the 128 countries included in both datasets.

The Democracy Index (DI) is calculated using a weighted average of a 60-item measure with items distributed over five categories: electoral process, civil liberties, government, political participation, and political culture. In addition to assigning a numeric DI to each of 167 countries, the EIU also classifies each country into one of our regime types by its index. The four regime types are (DI ranges in parentheses): Full Democracies (DI ≥ 8), Flawed Democracies (6 ≤ DI < 8), Hybrid Regimes (4 ≤ DI < 6), and Authoritarian Regimes (DI < 4).

The Energy Sustainability Index is a composite of three subscales:

- Energy security - the effective management of primary energy supply from domestic and external sources, the reliability of energy infrastructure, and the ability of participating energy companies to meet current and future demand.
- Energy equity - the accessibility and affordability of energy supply across the population.

- Environmental sustainability - the achievement of supply and demand-side energy efficiencies and the development of energy supply from renewable and other low-carbon sources.

We analyse both the overall sustainability index as well as the environmental impact mitigation subscale since it is the most directly relevant to our question of interest.

Consistent with our experimental results, we find that there is a significant positive correlation between democratic institutions and sustainability (Figure 4.6): countries with higher democracy scores also score higher in their efforts to act sustainably (Table 4.6 Col 1; $p < 0.001$) and specifically to mitigate environmental impact (Table 4.7 Col 1; $p < 0.001$).

As a first step towards testing the robustness of this relationship, we examine the effect of including controls for the 2014 gross domestic product (GDP, in US$, compiled by (World Bank 2013)), Gini index (a measure of wealth inequality, using the most recent year available for each country from (Quandl.com 2014)), literacy rates (using the most recent year available for each country, complied in the (CIA World Factbook 2014)), average life expectancy in each country in 2013 (complied by the (World Health Organization 2013)), the level of corruption in each country (compiled in 2004 by the (World Bank 2004)), each country's population size and the rate of unemployment (both using the most recent year available for each country from (Quandl.com 2014)). Missing values for controls are interpolated, using the mean of all non-missing values. As shown in Tables 5.6 col 2 and 5.7 col 2, we continue to find a significant positive relationship between democracy scores and both the overall sustainability index ($p = 0.001$) and the environmental impact mitigation score ($p < 0.001$). Finally, we also consider the effects of the logarithmic transformations of GDP and population size of each country, as these variables are heavily right-skewed. We find that our results remain unchanged (col 3 in Tables 5.6 and 5.7): more democratic institutions have more sustainable energy policies (p < 0.001) and engage in greater efforts to mitigate environmental impact (p < 0.001).

Thus we provide preliminary evidence that democracy may indeed lead to better environmental practices. Obviously this analysis does not control many additional potential confounding factors, and is correlational, not causal. We hope that these preliminary results will inspire empirical scholars to investigate this issue further in future work.

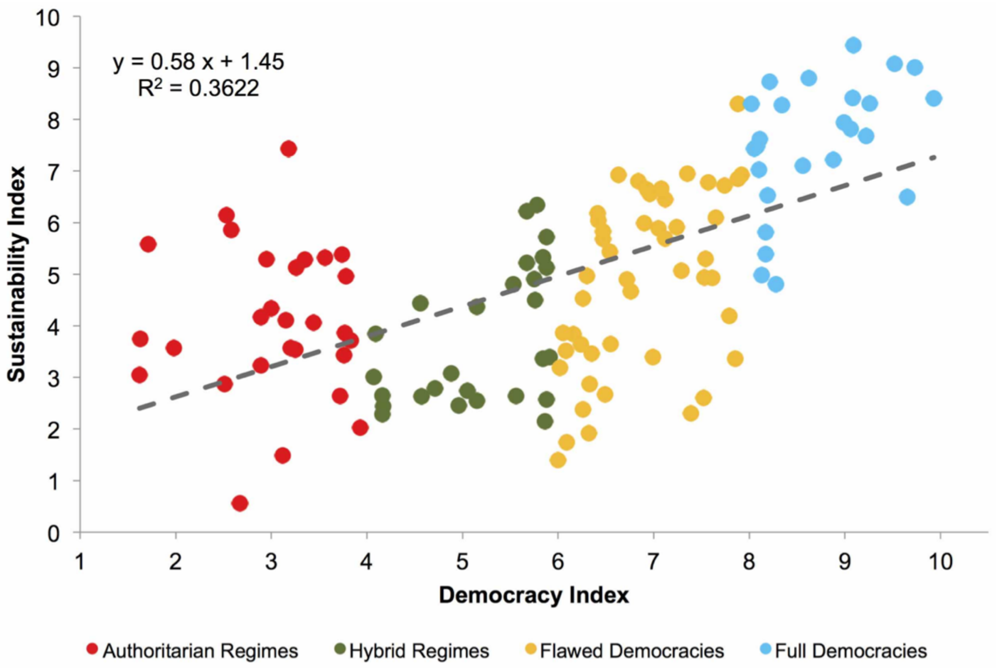

Figure 4.6. Real-world data support our experimental conclusions, as countries with more democratic governments have more sustainable energy policies. *Energy sustainability index (as measured by World Energy) is shown as a function of the Democracy Index (as measured by the Economist's Intelligence Unit) for N=128 countries. A strong positive association is clearly visible, and this association is robust to controlling for gross domestic product (GDP), Gini index, population size, literacy rate, unemployment rate, life expectancy, and level of corruption. Thus we provide preliminary empirical support for the role of democracy in promoting sustainability outside the laboratory. We adopt the colouring and naming scheme from the Economist Intelligence Unit's classification of regimes.*

Table 4.6. Linear regression predicting the overall sustainability index with democracy score.

	Sustainability Index	Sustainability Index	Sustainability Index
Democracy index	0.585 (0.073)***	0.211 (0.062)***	0.227 (0.054)***
GDP (in US$)		1.710e-07 (7.143e-08)*	
Gini index		-0.008 (0.013)	-0.004 (0.012)
Population size		-1.376e-06 (4.761e-07)**	
Literacy rate		2.753 (1.096)*	1.003 (0.963)
Unemployment rate		-2.651 (1.982)	-1.298 (1.703)
Life expectancy (years)		0.060 (0.028)*	0.006 (0.022)
Level of corruption		0.022 (0.006)***	0.011 (0.006)
log(GDP)			0.782 (0.170)***
log(Population)			-0.591 (0.181)**
Constant	1.454 (0.474)**	-3.703 (1.807)*	-1.424 (1.763)
R^2	0.36	0.63	0.72
N	128	128	128

* *p*<0.05; ** *p*<0.01; *** *p*<0.001

Table 4.7. Linear regression predicting the environmental impact mitigation score with democracy score.

	Environmental Impact Mitigation Score	Environmental Impact Mitigation Score	Environmental Impact Mitigation Score
Democracy index	0.652	0.662	0.668
	(0.107)***	(0.130)***	(0.130)***
GDP (in US$)		1.274e-08	
		(1.373e-07)	
Gini index		0.040	0.039
		(0.031)	(0.031)
Population size		-3.523e-06	
		(1.128e-06)**	
Literacy rate		-2.089	-2.022
		(2.077)	(2.295)
Unemployment rate		-1.598	-1.524
		(4.408)	(4.667)
Life expectancy (years)		0.030	0.038
		(0.049)	(0.056)
Level of corruption		0.002	0.005
		(0.013)	(0.015)
log(GDP)			-0.118
			(0.332)
log(Population)			-0.088
			(0.351)
Constant	1.029	-0.796	0.443
	(0.676)	(3.026)	(4.371)
R^2	0.21	0.26	0.24
N	128	128	128

* $p<0.05$; ** $p<0.01$; *** $p<0.001$

Appendix A.

Heterogeneity in background fitness acts as a suppressor of selection

A.1 Introduction

Evolutionary dynamics explores how strategies change over time and space in structured or unstructured populations (Bürger 2000; Durrett and Levin 1994; Helbing 2010; Fu, Liu, and Wang 2007; Hofbauer and Sigmund 2003; Imhof and Nowak 2006; Smith 1993; Nowak 2006a; Nowak and Sigmund 1992; Nowak and Sigmund 2004; Traulsen and Nowak 2006; Weibull 1997). These strategies can be alleles in a genetic context or behaviours in social interactions (Nowak, Tarnita, and Wilson 2010; Tarnita, Taubes, and Nowak 2013). In the simplest case, these strategies have a fixed fitness. Even in this case, population structure can have subtle influences, suppressing or amplifying selection (Allen and Tarnita 2014; Bürger 2000; Helbing 2010; Hofbauer and Sigmund 2003; Imhof and Nowak 2006; Lieberman, Hauert, and Nowak 2005; Nowak 2006a; Nowak and Sigmund 2004; Ohtsuki et al. 2006; Tarnita et al. 2009; Tarnita, Wage, and Nowak 2011; Traulsen, Claussen, and Hauert 2005; Weibull 1997). One important aspect of many real-world population structures is that different physical locations or positions in society have different value (Nowak, 2012): A good breeding site may give a breeding bird an advantage that is sometimes connected to its own behaviour (Kokko 1999) but sometimes also independent of its own behaviour (Misenhelter and Rotenberry 2000). A good school district can be influential for one's career progression (Cullen, Jacob, and Levitt 2005). Inherited wealth may positively affect reproductive success (Essock-Vitale 1984). We consider evolutionary dynamics in such a setting and ask how heterogeneity in the implicit value of different physical or societal positions affects the evolutionary dynamics. Our model does not include explicit spatial structure, but only considers different values for each position. In a biological context, this would mean that nesting site quality can crucially contribute to the spreading of new mutations, in addition to behavioural or physiological change associated with this mutation. In a social interpretation, it would mean that we imitate successful individuals, assuming their success derives from a behaviour we might be able to copy. It may also be the case, however, that we not only imitate those who are successful due to their behaviour, but also those who are successful due to heritage or their social and economic ties. In the latter case, the imitation may be in vain, but this does not preclude strategies from spreading.

In our approach, we assume that fitness is the sum of the background fitness associated with a certain position (or location) and the fitness derived from the strategy of an individual. We assume that a strategy can spread from any position to any other position through individuals copying each other. Thus, we can use the convenient mathematical properties of well-mixed, unstructured populations when it comes to the changes in the abundance of a strategy. At the same time, however, the distribution in background fitness allows to address an important aspect of population structure that has not been considered in this context so far. In contrast, spatial and social heterogeneity has been considered in the case of evolutionary dynamics in degree-heterogeneous networks (Lieberman, Hauert, and Nowak 2005; Ohtsuki et al. 2006; Perc and Szolnoki 2008; Poncela et al. 2009; F. C. Santos and Pacheco 2005; Pacheco et al. 2011; F. C. Santos et al. 2012; F. C. Santos, Santos, and Pacheco 2008; Szabó and Fath 2007). Another source of heterogeneity arises from different kinds of interactions within the population (Chatterjee, Zufferey, and Nowak 2012; Fu et al. 2008; McNamara, Barta, and Houston 2004; Rand, Tarnita, and Ohtsuki 2013; Taylor and Nowak 2006; Traulsen, Pacheco, and Nowak 2007; J. Wang, Fu, and Wang 2010). Also in population genetics, heterogeneity in offspring number and nest sites has been addressed (Eldon and Wakeley 2006; Lessard 2007; Wakeley 2008).

Our model is based on a Markov chain with two absorbing states – a new strategy is eventually either lost or reaches fixation in a finite population. In homogeneous populations, the transition matrix of these processes of reduces to a tri-diagonal matrix, leading to closed expressions for the time to absorption or the probability to reach a certain state (Altrock and Traulsen 2009; Nowak et al. 2004). In the case of heterogeneous wealth distribution, such an approach fails and these quantities typically must be inferred numerically based on standard methods (Grinstead and Snell 2012). However, the same method leads to a full analytical solution in closed form for small populations. For larger populations, the corresponding analytical expressions become cumbersome, but a Taylor expansion of the small population result in the important limit of large heterogeneity gives us an approximation that numerically also holds for larger populations. Throughout this paper, we adopt terms (e.g. "wealth", "rich", "poor", "inequality" etc.) inspired by economics and sociology. But a biological meaning for each of these words can readily be inferred (e.g. "resources", "high quality of nest site", "low quality of nest site", "heterogeneous nest site qualities" etc.).

A.2 An evolutionary process with heterogeneous background fitness

We assume a finite population of size *N* with two types A and B. Evolution proceeds by selecting one individual proportional to its total fitness to reproduce asexually. Its identical offspring replaces another individual chosen with uniform probability to die (Moran 1962). This implies that in each time step, the number of individuals of a certain type can change at most by ± 1. Hence, the dynamics can be captured by a simple birth-death process, which allows calculating the probability of fixation and the associated time as well as several related quantities in closed form analytically (Antal and Scheuring 2006; Nowak et al. 2004). When mutations arise infrequently, the fixation probability is a relevant measure to describe the average abundance of types in a mutation-selection equilibrium (Fudenberg and Imhof 2006). In this case, a mutant will fixate or go extinct before another mutant arises. Thus, the system effectively reduces to an evolutionary process jumping between the two absorbing states where all individuals use the same strategy.

An individual's total fitness *f* is the sum of that individual's background fitness b and the fitness derived from the strategy *s* the individual has chosen:

$$f_i = b_i + s_i \tag{1}$$

where i $(0 \leq i \leq N)$ denotes an individual in the population. Note that we assume that the strategy of the individual has an impact on the fitness that is only dependent on the individual's type. We assume no frequency-dependent interactions between types, such that $s_i > 0$ is a fixed number. Due to heterogeneities in the background fitness b_i, however, our state space is not only determined by the number of individuals of one type, but also by the unique position of each individual. Therefore the transition matrix is no longer tri-diagonal, excluding many analytical approaches based on this property. Thus, calculating a closed form for the absorption probabilities and times becomes much more cumbersome.

We assume that an offspring inherits its parent's strategy, but it does not receive its parent's background fitness. Instead, the offspring "inherits" the background fitness of the individual who was chosen for death and thus previously occupied the same location. In other words, the topology of background fitness remains unchanged over time, but strategies evolve on top of the background fitness topology. The fixed background topology thus represents a static environment in which the strategies change due to biological or cultural reproduction. Such an environment could be breeding sites in biology (Misenhelter and Rotenberry 2000) or economic wealth in human society (Wolff 2002).

A.3 Background fitness effectively reduces intensity of selection

We assume there exist two strategies A and B. If $s_A > s_B$ there is constant selection for type A and if $s_A < s_B$ selection favours B. Thus $s_A = s_B$ is the neutral case. Without loss of generality, we assume that strategy B's fitness is always $s_B = 1$. All values of strategy fitness and background fitness are non-negative.

We are interested in the fixation probability of a single mutant of type A in a population of $N - 1$ individuals of type B. Let ρ_i and τ_i denote the fixation probability and average absorption time of type A if the mutant arises in location i, and let ρ and τ denote the average fixation probability and absorption time of type A if the mutation arises at a random location in the population:

$$\rho = \frac{1}{N}\sum_{i=1}^{N} \rho_i \qquad (2)$$

$$\tau = \frac{1}{N}\sum_{i=1}^{N} \tau_i \qquad (3)$$

We combine an analytical approach, which is feasible for small populations only, with computational approaches. Numerically, we compute properties from the exact transition matrix of the Markov Chain and run stochastic agent-based simulations. Agent-based simulations proceed as follows: in every time step, one individual is selected proportional to fitness to reproduce and one individual is selected at random to die, until the population has reached a homogeneous state in which all individual are of type A or B. We average over m realisations for every possible initial location of the mutant to calculate ρ_i. Thus, the fixation probability of a randomly arising mutant ρ is the average over Nm realisations.

We are interested in the effect that heterogeneous background fitness has on the fixation probability ρ of a randomly arising mutant. We expect that heterogeneity in background fitness modulates and neutralises the effects of selection on a strategy, similar to some population structures (Lieberman, Hauert, and Nowak 2005; Traulsen, Claussen, and Hauert 2005) or the introduction of a random number of interactions in evolutionary games (Traulsen, Nowak, and Pacheco 2007). Intuitively, the strength of selection becomes effectively weaker when any background fitness is introduced. This is because the total fitness of an individual is no longer just derived from using a strategy, but also from a fixed, non-negative background fitness (see Equation (1)). Depending on their relative value, either strategy-dependent fitness or individual background fitness may have a greater impact on the total fitness.

Second, if there exists any heterogeneity in background fitness, richer individuals will be favoured over poorer individuals and thus mutants arising in rich individuals are more likely to spread. This is especially important if background fitness values are large compared to the strategy payoff values. Moreover, the effect of heterogeneity on fixation ought to be largest when the total background fitness in the population is very unequally distributed among individuals of a population. We speak of *perfect inequality* when one individual j in a population possesses the entire wealth $K = \sum_{i=1}^{N} b_i$ in the population (Keister and Moller 2000): $b_j = K$ and $b_i = 0$ for all remaining individuals.

An intuitive prediction is that for a large enough value of total background fitness K in the population and under perfect inequality, the probability of fixation approaches the case of neutral drift, $1/N$. In other words, if the total amount of background fitness is large and all of it is in the possession of one individual j, then fixation of type A occurs *only if* the mutant arises in individual j. Mutations arise at a random position in the population, and therefore the probability of fixation must converge to $1/N$.

We numerically confirm this hypothesis of perfect inequality (Figure A.1): We vary the fitness r of the mutant type A in a population of size $N = 30$ to find the fixation probability for 4 values of the rich individual's wealth, b_1. For $b_1 = 0$, we recover the well known results for constant selection of a mutant with fitness r in a *homogeneous* population. As b_1 increases, the fixation probability increases for disadvantageous mutants, $r < 1$, and decreases for advantageous mutants, $r > 1$. In other words, by increasing the amount the rich individual possesses, fixation of the randomly arising mutant strategy becomes less dependent on the strategy's advantage or disadvantage and more dependent on the origin of the mutation: The quality of an individual's position, not its strategy, becomes the determinant for evolutionary success. For $b_1 = 100$, the probability of fixation is already very close to the neutral case where $s_1 = s_2$. This is an example of how high inequality in background wealth suppresses selection.

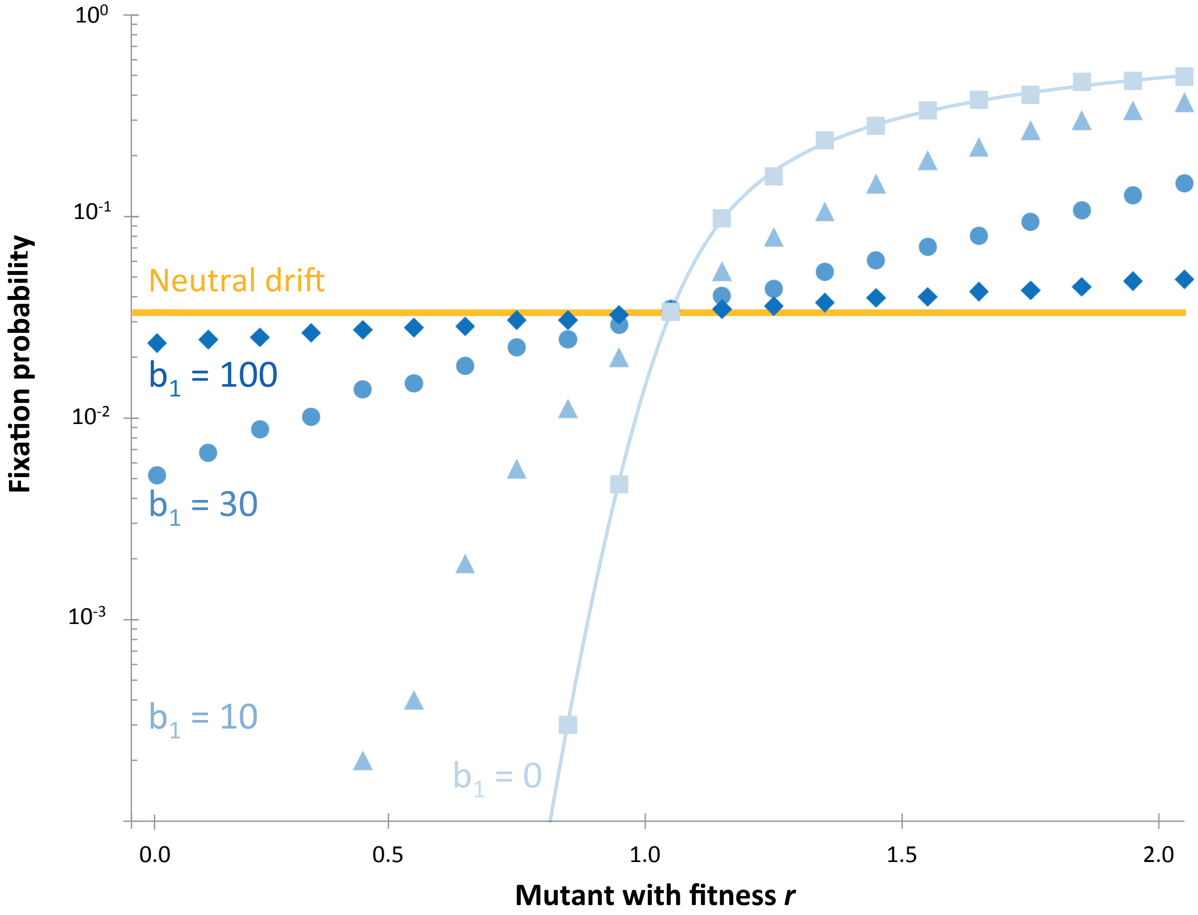

Figure A.1. Constant selection is suppressed as inequality in background fitness increases. *A mutant with fitness r arises in a population of size* $N = 30$*. The population is heterogeneous: b_1 denotes the background fitness of one individual, while all others have background fitness 0. For* $b_1 = 0$*, we recover the well known fixation probability under constant selection,* $\frac{1-\frac{1}{r}}{1-\frac{1}{r^N}}$ *(solid line). For very high background fitness and large inequality,* $b_1 \gg 1$*, the fixation probability of a randomly arising mutant approaches* $1/N$*. Thus, selection is suppressed through the introduction of heterogeneous background fitness (symbols denote individual-based simulations averaged over 30,000 realisations).*

A.4 Unequal wealth distributions suppress selection more than an equal distribution

We note that in Figure A.1, as the wealth of the only rich individual b_1 increases, so does the total wealth K. In other words, the total wealth in the population is increasing but all of it is in the possession of one individual. All of the examples with $b_1 > 1$ in Figure 3.1 are thus examples of perfect inequality, although the extent to which the rich individual possesses more background wealth varies greatly. But this raises the question if selection is suppressed because the total wealth has increased or because inequality has increased.

We therefore separate the effects of K and b_1 on fixation. To this end, we are not only interested in the fixation probability of type A when $b_1 = K$ but also when wealth is distributed differently in the population. Wealth can be distributed in many different ways within the population. We focus on those distributions where the total wealth is split by an increasing number of individuals within the population. That is, in the most unequal case, the total wealth is in the hands of one individual ($j = 1$). The same amount of total wealth is then successively owned by two players ($j = 2$), then by three ($j = 3$), and so on, until it is split evenly among all individuals ($j = 15$) (Figure A.2).

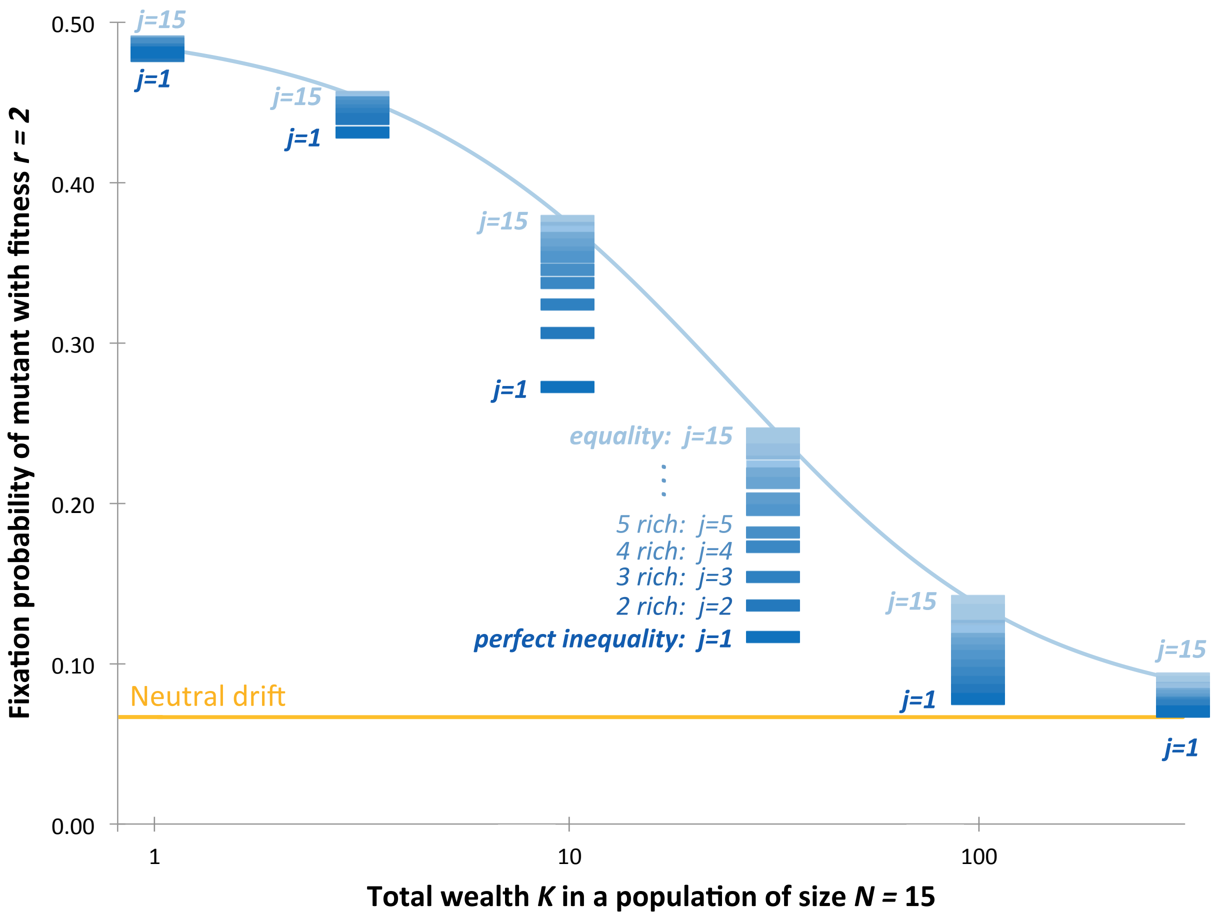

Figure A.2. Increasing a population's total wealth suppresses selection. Inequality amplifies the effect that wealth has on the fixation probability. *As the total wealth in a population of* $N = 15$ *increases, the fixation probabilities of any wealth distribution shift towards neutral drift. A wealth distribution here refers to the number of rich actors j that share the total wealth in the population. The fewer rich people own the total population wealth (the smaller j), the more selection is suppressed. Perfect inequality* ($j = 1$) *is the strongest suppressor of selection, regardless of the total wealth* K. *In contrast, when all individuals are equally wealthy, the fixation probability is least suppressed for any given total wealth in the population* ($j = 15$). *Lines show the fixation probabilities in a homogeneous population and neutral drift.*

We clearly observe that both forces are driving the suppression of selection (Figure A.2): the total amount of wealth in the population and also the inequality in the distribution of the total wealth within the population affect the effective intensity of selection. An increase in the total wealth in the population also leads to the probability of any fixation being closer to neutrality: it defines the possible magnitude to which selection can be suppressed at a minimum and at a maximum. How much selection is suppressed is then entirely determined by the inequality in wealth distribution. The more unequal wealth is distributed, the more selection is suppressed (i.e. closer to neutrality within the possible magnitude defined by the total wealth). Conversely, the more equally wealthy everyone is, the less the effect of selection suppression is.

If there exists no background wealth, heterogeneity is obviously absent and has no effect. As background wealth increases, inequality suppresses selection at a faster rate than does equality of the same total wealth K, which leads to the observed disparity of fixation probabilities in the middle section of K (Figure A.2). This trend, however, is bound by a global minimum (that is, neutral drift) for both inequality and equality. Hence, as total background wealth becomes very large, selection is suppressed to a similar extend by both inequality and equality because neither can suppress selection below $1/N$.

These results show that selection is suppressed when background wealth is added. Intuitively, this makes sense because with higher background fitness, the addition of a small fitness value from the strategy has little impact. This effect is amplified by an unequal allocation of wealth within the population. To understand the exact mechanism, we use an analytical approach to study heterogeneity. We analytically solve the fixation probabilities in populations of size $N = 3, 4$ and 5. We also provide an approximation that numerically holds in the limiting case for larger population sizes.

A.5 Perfect inequality is a strong suppressor of selection

Based on the Markov Chain process, we construct a stochastic transition matrix $\mathcal{M}$ such that $\mathcal{M}_{jk}$ is the transition probability to go from j to k; $j, k \in \{1, \dots, 2^N\}$. The state space of the transition matrix is described by all possible binary strings of 1s (A) and 0s (B). The absorption probabilities ρ_i into A are the elements of the eigenvector to the largest eigenvalue of $\mathcal{M}$ (Grinstead and Snell 2012). We find the fixation probability ρ and absorption time τ of a randomly arising mutant numerically based on Equations 2 and 3.

The transition matrix is of size $2^N \times 2^N$. Therefore closed analytical solutions of our model are only feasible for small N. We use numerical solutions based on the transition matrix because, in contrast to stochastic agent-based simulations, solutions derived directly from the Markov Chain process are exact and do not require a large number of realisations. For an analytical solution, which leads to results that can be interpreted most easily, we rearrange the matrix elements in $\mathcal{M}$ into its canonical form and derive the fundamental matrix containing information on all transient states (Grinstead and Snell 2012). In the canonical form, states are renumbered such that the p transient states come first, followed by the two absorbing states where all individuals use the same strategy.

$$\mathcal{M} = \begin{pmatrix} \mathcal{Q} & \mathcal{R} \\ \mathcal{O} & \mathcal{J} \end{pmatrix} \tag{4}$$

where $\mathcal{J}$ is a 2×2 identity matrix (once all individuals use the same strategy, the state is not left again), $\mathcal{O}$ is a $p \times 2$ zero matrix (one cannot escape from an absorbing state), $\mathcal{R}$ is a non-zero $2 \times p$ matrix describing fixation, and $\mathcal{Q}$ is a $p \times p$ matrix describing the dynamics within the transient states. Here $p = 2^{N-1}$ is the number of unique positions of all individuals in a population of size N.

For an absorbing Matrix Chain, there exists an inverse $\mathcal{N}$ of the matrix $\mathcal{I} - \mathcal{Q}$. $\mathcal{N}$ is called the fundamental matrix of $\mathcal{M}$, and $\mathcal{N} = \mathcal{I} + \mathcal{Q} + \mathcal{Q}^2 + \cdots$ (Grinstead and Snell 2012). Each entry $\mathcal{N}_{ij}$ of $\mathcal{N}$ contains the expected number of time steps the chain is in state j, given that it starts in state i. Hence, the time to absorption τ_i, given that the process starts in state i, is the sum over all entries of $\mathcal{N}$ in row i. Let γ_{ij} be the probability that the process will be absorbed in the absorbing state j, given that it starts in the transient state i. Let γ be the $p\times 2$ matrix with entries γ_{ij}:

$$\gamma = \mathcal{N}\mathcal{R} \tag{5}$$

Specifically, we speak of a fixation probability ρ if the process begins in a transient state i where only 1 mutant exists. This is only the case for a subset of entries in γ. The size of this subset depends only on the distribution of the background fitness.

We can perform the above procedure analytically and thus obtain a closed solution of our model in the case of small population sizes $N = 3, 4$ and 5. While analytical solutions for larger population sizes are theoretically attainable, they come in intricate form and are very difficult to interpret beyond $N = 5$. This is because the starting point for an analytical solution is the transition matrix and the analytical procedure implies that we need to find an analytical form for its eigenvector. As noted above, the transition matrix is of size $2^N \times 2^N$. Therefore, this procedure quickly becomes very cumbersome as N increases.

We calculate the exact solutions for the fixation probabilities ρ for $N = 3$ for any arbitrary background wealth distribution $\mathcal{K} = (b_1, b_2, b_3)$ from Equation (5). Despite the small population size, the exact general solution for an arbitrary wealth distribution spans multiple pages.[1] For ease of reading, we therefore only print the solution in the case of perfect inequality and simplify our notation: we set $b_1 = b$ for the rich individual and $b_2 = b_3 = 0$ for the two poor individuals. We then obtain for the fixation probability of the rich individual:

$$\rho_{rich} = \frac{(1+b)(r^3\left(7+7b+2b^2\right)+2r^4(2+b)+(1+b)(2+b)^2)}{2(2+3b+b^2)+r(11+16b+9b^2+2b^3)+r^2(15+22b+16b^2+6b^3+b^4)+r^3(11+16b+9b^2+2b^3)+2r^4(2+3b+b^2)} \tag{6}$$

Similarly, for each of the two poor individuals, we find:

$$\rho_{poor} = \frac{r^2\left(4+3b+b^2\right)+r^3(7+8b+4b^2+b^3)+2r^4(2+3b+b^2)}{2(2+3b+b^2)+r(11+16b+9b^2+2b^3)+r^2(15+22b+16b^2+6b^3+b^4)+r^3(11+16b+9b^2+2b^3)+2r^4(2+3b+b^2)} \tag{7}$$

Finally, using Equation 2, we can also calculate the fixation probability of a randomly arising mutant $\rho = \frac{1}{3}\left(\rho_{rich} + 2\rho_{poor}\right)$:

$$\rho = \frac{r^2(12+18b+15b^2+6b^3+b^4)+r^3(21+30b+17b^2+4b^3)+6r^4(2+3b+b^2)}{3(2(2+3b+b^2)+r(11+16b+9b^2+2b^3)+r^2(15+22b+16b^2+6b^3+b^4)+r^3(11+16b+9b^2+2b^3)+2r^4(2+3b+b^2))} \tag{8}$$

When $b = 0$, no background wealth is present in the population and Equations (6), (7) and (8) reduce to the well known fixation probability in the case of a homogeneous environment without any background wealth, $\frac{1-\frac{1}{r}}{1-\frac{1}{r^3}}$.

[1] A complete solution can be obtained from ohauser@fas.harvard.edu.

Furthermore, the analytical solution approaches $1/N$ in the limit of $b \to \infty$ (Figure A.3a), which is also expected from our argument above: When the total wealth in the population is large and one individual owns all this wealth, fixation can only occur if the mutation arises in the position with high background fitness (Figure A.3b).

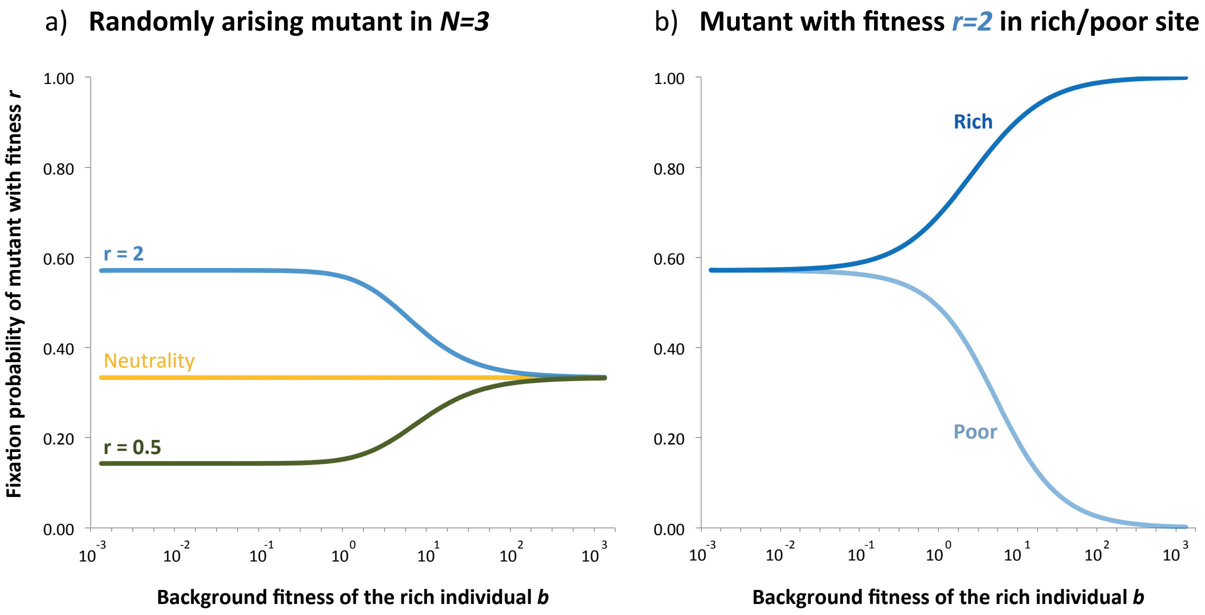

Figure A.3. In a population of size $N = 3$, the fixation probability of a randomly arising mutant approaches $1/3$ as inequality increases because the mutation can only fixate if it arises in the wealthy individual. *a) For disadvantageous ($r = 0.5$) and advantageous mutant ($r = 2$), selection is suppressed and approaches neutrality $1/N = 1/3$ for high inequality, $b \gg 1$. b) As inequality increases, the probability of fixation of mutant $r = 2$ increases for the wealthy individual and decreases for all others. This argument holds for any value of r.*

The exact solutions from the Markov Chain process provide very useful insight into the evolutionary dynamics in a heterogeneous population. They are, however, long and complicated to interpret. We therefore also present approximated solutions that are more intuitive and easily derived from previous results. For each of the three population sizes $N = 3, 4$ and 5, we find the first-order Taylor approximations for large b for the equations corresponding to (6), (7), and (8):

$$\rho_{rich} \approx 1 - \frac{N-1}{rb} \tag{9}$$

$$\rho_{poor} \approx \frac{r}{b} \tag{10}$$

$$\rho \approx \frac{1}{N} + \frac{N-1}{N}\frac{r}{b}\left(1 - \frac{1}{r^2}\right) \tag{11}$$

For very large b, only the leading term in Equation (11) prevails, $\rho \approx 1/N$, which is in agreement with the limiting case discussed above. We find that the first order approximations work very well for small populations with high population wealth under perfect inequality (for instance, the fixation probability of a randomly arising mutant using Equation (11) for $N = 3$ is shown in Figure A.4a). Moreover, numerical simulations suggest that the Taylor approximations are also in reasonable agreement with slightly larger population sizes (Figure A.4b,c). In general, we expect this approximation to hold for larger N if the wealth of the rich individual also increases. That is, $1/b$ enters in a similar way as r does in the weak selection approximation of the fixation probability under constant selection, $\rho = \frac{1-\frac{1}{r}}{1-\frac{1}{r^N}}$. In that case, the approximation is valid only for $N(r-1) \ll 1$. Similarly, in the present case the convergence radius also depends on N, b and r. For example, in the case of large r, $(N-3)r^2/((1+r)b) \ll 1$ is required.

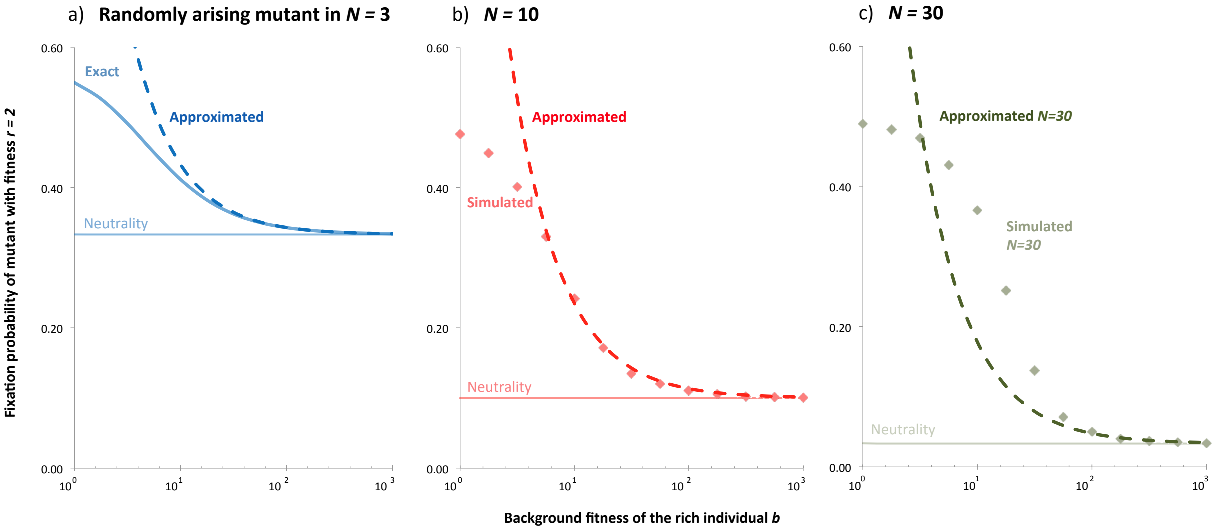

Figure A.4. If inequality and background fitness are large, a simple approximation of the rate of evolution is derived from the analytical solution for small N. Numerically, it is also reasonable for larger N. *a) For a randomly arising mutant in* $N = 3$*, the fixation probability can be approximated as* $\rho \approx \frac{1}{N} + \frac{N-1}{N}\frac{r}{b}\left(1 - \frac{1}{r^2}\right)$ *for large b. b) Although the approximation is analytically derived only for population sizes* $N \leq 5$*, it also appears to hold for slightly larger* N*, such as* $N = 10$ *and, for very large background wealth,* $N = 30$*. Legend: solid lines are exact (Markov Chain) solutions, dashed lines are first order Taylor expansions of the solutions for large inequality and symbols represent agent-based simulations over 30,000 realisations.*

A.6 Time to absorption decreases in a wealthy population under perfect inequality

In addition to the probability of fixation of a mutant, we are also interested in how fast the mutant either fixates or goes extinct. The absorption time measures how many time steps need to be taken on average to reach either of the two absorbing states – that is, either a state of all A or all B individuals. It can be calculated from summing over the line corresponding to the initial condition in the fundamental matrix $\mathcal{N}$ of the Markov process (Grinstead and Snell 2012).

In a population of size $N = 3$, the unconditional absorption times of a rich and poor individual are, respectively:

$$\tau_{rich} = \frac{3(4(2+3b+b^2)+r(30+52b+34b^2+8b^3)+r^2(48+80b+55b^2+18b^3+3b^4)+r^3(37+54b+27b^2+6b^3)+6r^4(2+3b+b^2))}{2(2(2+3b+b^2)+r(11+16b+9b^2+2b^3)+r^2(15+22b+16b^2+6b^3+b^4)+r^3(11+16b+9b^2+2b^3)+2r^4(2+3b+b^2))} \tag{12}$$

$$\tau_{poor} = \frac{3(4(2+3b+b^2)+2r(15+20b+9b^2+2b^3)+r^2(48+65b+41b^2+12b^3+2b^4)+r^3(37+54b+30b^2+7b^3)+6r^4(2+3b+b^2))}{2(2(2+3b+b^2)+r(11+16b+9b^2+2b^3)+r^2(15+22b+16b^2+6b^3+b^4)+r^3(11+16b+9b^2+2b^3)+2r^4(2+3b+b^2))} \tag{13}$$

The unconditional absorption time of a randomly arising mutant (Equation 3) thus is:

$$\tau = \frac{12(2+3b+b^2)+2r(45+66b+35b^2+8b^3)+r^2(144+210b+137b^2+42b^3+7b^4)+r^3(111+162b+87b^2+20b^3)+18r^4(2+3b+b^2)}{2(2(2+3b+b^2)+r(11+16b+9b^2+2b^3)+r^2(15+22b+16b^2+6b^3+b^4)+r^3(11+16b+9b^2+2b^3)+2r^4(2+3b+b^2))} \tag{14}$$

Again, these three equations reduce to the well-known result from homogeneous populations for $b = 0$ (Altrock and Traulsen 2009).

When we look at the absorption time of a randomly arising mutant (Equation (14)), we find that higher inequality leads to fewer time steps until absorption (Figure A.5a), relative to the same amount of wealth being equally distributed among all individuals. When we look at the individual components of the absorption time, we find that mutations arising in poor individuals generally absorb more quickly than equally wealthy individuals or rich individuals (Figure A.5b). Intuitively, the higher the inequality in a population, the quicker the poor individuals absorb back into the resident state. This is a consequence of inequality as it hinders poor individuals from fixating and fosters their extinction.

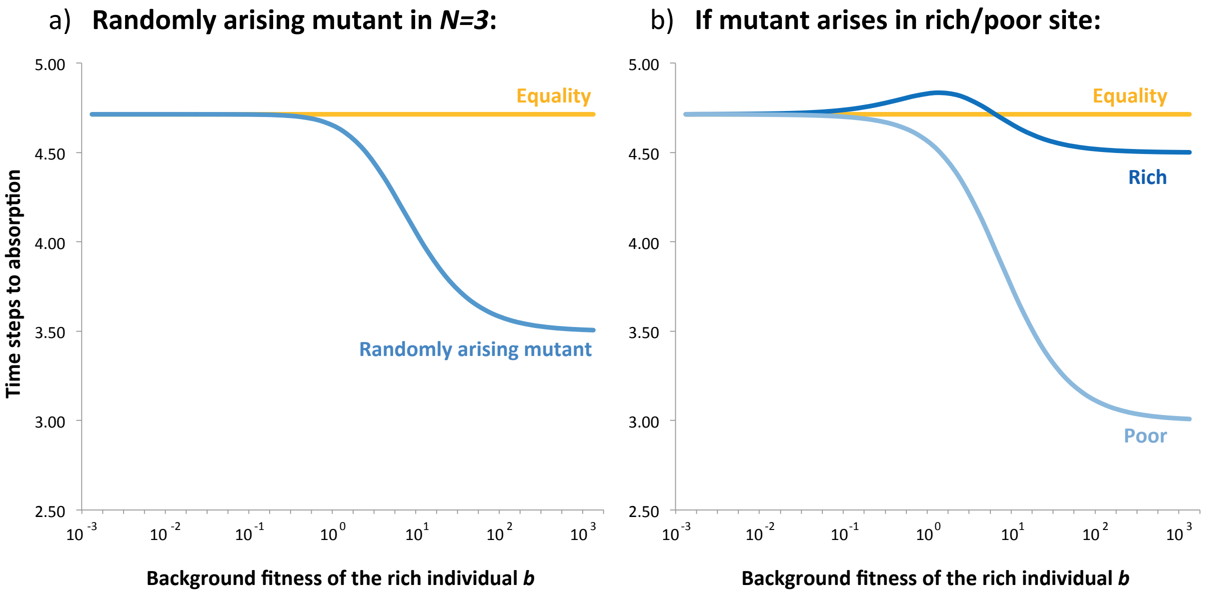

Figure A.5. The absorption time of a randomly arising mutant decreases with inequality. *a) In contrast to a population with equal wealth distribution, inequality reduces the number of time steps that a randomly arising mutant in a population of* $N = 3$ *makes on average to absorption. b) This effect in reduction of absorption time is driven primarily by the fast extinction of poor individuals when inequality disfavours them. In contrast, if inequality exists but it is small, richer individuals take longer to fixate than if all were equal. This is because the advantage from background wealth is small and rich individuals are therefore only selected slightly more often for reproduction than poor individuals.*

The absorption time of the rich individual depends on the relative advantage that the background wealth bestows on it. When the background wealth of the rich individual is comparable to the payoff values of the competing strategies, the time to absorption is longer. This is because the advantage from background wealth is small and rich individuals are therefore only selected slightly more often for reproduction than poor individuals.

As the wealth advantage increases, however, the time to absorption decreases rapidly. The advantage that the background wealth provides to the rich individual increases and selection favours this individual for reproduction. The rich individual's strategies reach fixation more often and more quickly than the poor individuals' strategies.

A.7 Discussion

We have shown that heterogeneity in background fitness reduces selection regardless of the strategy payoff. Background fitness is a concept that has been previously proposed to incorporate environmental or otherwise contributing factors to competitiveness, fitness and evolutionary survival (Dayton-Johnson and Bardhan 2002; Deng, Tang, and Zhang 2011; McNamara, Barta, and Houston 2004). Equal background reduces the intensity of selection: the payoff values of each strategy are scaled proportionally to the homogeneous background fitness of all individuals (Deng, Tang, and Zhang 2011; Nowak et al. 2004). We focus on shifting the distribution of background fitness among individuals in a population. A heterogeneous distribution of background fitness is a possible way to address this issue. Such heterogeneities may exist in economic, social, cultural, or other dimensions that are all prevalent in nature and society (Dayton-Johnson and Bardhan 2002; Droz, Szwabiński, and Szabó 2009; Norton and Ariely 2011; Perc and Szolnoki 2008; F. C. Santos, Santos, and Pacheco 2008).

Our results show that inequalities in background fitness can lead to suppressed selection. Many types of suppressors of selection are known (Antal and Scheuring 2006; Lieberman, Hauert, and Nowak 2005; Nowak, Michor, and Iwasa 2003; Traulsen, Claussen, and Hauert 2005). The largest extent to which selection can be suppressed is if the affected trait fixates at random in a population, neutralising the effects derived from fitness entirely (Nowak, Michor, and Iwasa

2003). We show that large inequality can lead to this kind of neutral drift. Our work complements an existing literature in population genetics on heritable traits when selection is weak and elements of population structure are heterogeneous (Eldon and Wakeley 2006; Lessard 2007).

The evolution of frequency-dependent traits, such as cooperation, under heterogeneous background fitness is another interesting aspect alongside constant selection for a heritable trait. It has been argued that inequality can either increase cooperation in the public goods management (Olson 1965; J. Wang, Fu, and Wang 2010) or, much in contrast, lead to the downfall of cooperation (Varughese and Ostrom 2001). Neither effect has been discussed in the context of evolutionary biology. Our work provides a first step towards (constant) evolutionary games in a finite population, by separating out the effects of interaction from the background fitness of individuals.

Moreover, evolutionary game theory has received much attention on networks (Abramson and Kuperman 2001; Fu et al. 2008; Lieberman, Hauert, and Nowak 2005; Perc and Szolnoki 2010; F. C. Santos and Pacheco 2005; Szabó and Fath 2007). Many networks have been studied in regards to imitation of traits or spread of pathogens or emotions (Christakis and Fowler 2007; Christakis and Fowler 2008; Hill et al. 2010). Recently the interest in directed networks (Masuda and Ohtsuki 2009) or degree-heterogeneous networks has increased (Antal and Scheuring 2006; F. C. Santos, Pacheco, and Lenaerts 2006) and produced stimulating results. It would be interesting finding the connection between heterogeneous background fitness and heterogeneous networks.

In our model, we have also shown that absorption times under constant selection are negatively affected by heterogeneous background fitness. In other words, the more inequality exists, the faster absorption takes place. This stands in contrast to findings on a graph with heterogeneous edge weights in which the mean absorption time increases (Voelkl 2010). While heterogeneity leads to an increase of absorption time in some models (Frean, Rainey, and Traulsen 2013), it can be a catalyst to determine whether or not a strategy goes to fixation. We find that, when fixation time is very fast in an unequal population, often the mutant went extinct after arising in a poor individual. This is the case because a poor individual is unlikely to be selected for reproduction when inequality in background fitness is large.

Finally, the co-evolution of background fitness and strategic fitness could lead to interesting dynamics. The tendency that individuals in different classes of background fitness might show towards choosing an appropriate strategy is a crucial feature of many real-world examples, such as differences in votes between social classes over tax reforms (Ogburn and Peterson 1916). Other evidence comes from field studies (Dayton-Johnson and Bardhan 2002) that show that head-end and tail-end farmers in irrigation systems derive different incentives from their location, which in turn influences their strategy whether or not to cooperate.

Bibliography

Abeler, Johannes, Armin Falk, Lorenz Goette, and David Huffman. 2011. “Reference Points and Effort Provision.” *American Economic Review* 101 (2): 470–92. doi:10.2307/29783680.

Abramson, Guillermo, and Marcelo Kuperman. 2001. “Social Games in a Social Network.” *Physical Review E* 63 (3). APS: 030901.

Alesina, A, and G-M Angeletos. 2005. “Fairness and Redistribution” 151 (3712): 867–68. doi:10.2307/4132701.

Allen, Benjamin, and Corina E Tarnita. 2014. “Measures of Success in a Class of Evolutionary Models with Fixed Population Size and Structure.” *Journal of Mathematical Biology* 68 (1-2). Springer: 109–43.

Allison, P D. 1978. “Measures of Inequality.” *American Sociological Review* 43 (6): 865. doi:10.2307/2094626.

Altrock, Philipp M, and Arne Traulsen. 2009. “Fixation Times in Evolutionary Games Under Weak Selection.” *New Journal of Physics* 11 (1). IOP Publishing: 013012.

Ambrus, Attila, and Ben Greiner. 2012. “Imperfect Public Monitoring with Costly Punishment: an Experimental Study.” *American Economic Review* 102 (7): 3317–32. doi:10.1257/aer.102.7.3317.

Amir, Ofra, David G Rand, Ya'akov Kobi Gal, and Ya'akov Kobi Gal. 2012. “Economic Games on the Internet: the Effect of $1 Stakes.” Edited by Matjaz Perc. *PLoS ONE* 7 (2): e31461. doi:10.1371/journal.pone.0031461.

Anderson, Lisa R, Jennifer M Mellor, and Jeffrey Milyo. 2008. “Inequality and Public Good Provision: an Experimental Analysis.” *The Journal of Socio-Economics* 37 (3): 1010–28. doi:10.1016/j.socec.2006.12.073.

Andreoni, James. 1988. “Why Free Ride?.” *Journal of Public Economics* 37 (3): 291–304. doi:10.1016/0047-2727(88)90043-6.

Antal, Tibor, and Istvan Scheuring. 2006. “Fixation of Strategies for an Evolutionary Game in Finite Populations.” *Bulletin of Mathematical Biology* 68 (8). Springer: 1923–44.

Antinyan, A, L Corazzini, and D Neururer. 2015. “Public Good Provision, Punishment, and the Endowment Origin: Experimental Evidence.” *Journal of Behavioral and ….*

Ariely, Dan. 2008. *Predictably Irrational: the Hidden Forces That Shape Our Decisions*. Harper Perennial.

Atkinson, A B, and Thomas Piketty. 2007. *Top Incomes Over the Twentieth Century: a Contrast Between European and English-Speaking Countries*. Oxford University Press.

Baland, Jean-Marie, and Jean-Philippe Platteau. 1997. "Wealth Inequality and Efficiency in the Commons, Part I: the Unregulated Case." *Oxford Economic Papers* 49 (4). Oxford University Press: 451–82.

Barrett, S, and A Dannenberg. 2012. "Climate Negotiations Under Scientific Uncertainty." In. doi:10.1073/pnas.1208417109/-/DCSupplemental/pnas.201208417SI.pdf.

Benkler, Yochai. 2011. *The Penguin and the Leviathan: How Cooperation Triumphs Over Self-Interest*. Random House.

Berinsky, A J, G A Huber, and G S Lenz. 2012. "Evaluating Online Labor Markets for Experimental Research: Amazon. Com's Mechanical Turk." *Political Analysis* 20: 351–68. doi:10.1093/polana/mpr057/-/DC1.

Bernard, Mark, Anna Dreber, Pontus Strimling, and Kimmo Eriksson. 2013. "The Subgroup Problem: When Can Binding Voting on Extractions From a Common Pool Resource Overcome the Tragedy of the Commons?." *Journal of Economic Behavior & Organization* 91 (July). Elsevier B.V.: 122–30. doi:10.1016/j.jebo.2013.04.009.

Bornstein, Gary, and Ori Weisel. 2010. "Punishment, Cooperation, and Cheater Detection in ``Noisy‘' Social Exchange." *Games* 1 (March): 18–33. doi:10.3390/g1010018.

Boyd, R, and P J Richerson. 1988. "The Evolution of Reciprocity in Sizable Groups." *Journal of Theoretical Biology* 132 (3): 337–56.

Brick, Kerri, and Martine Visser. 2012. "Heterogeneity and Voting: a Framed Public Good Experiment." *Working Paper*, July, 1–18.

Buckley, Edward, and Rachel Croson. 2006. "Income and Wealth Heterogeneity in the Voluntary Provision of Linear Public Goods." *Journal of Public Economics* 90 (4-5): 935–55. doi:10.1016/j.jpubeco.2005.06.002.

Buhrmester, M, T Kwang, and S D Gosling. 2011. "Amazon's Mechanical Turk: a New Source of Inexpensive, Yet High-Quality, Data?." *Perspectives on Psychological Science* 6 (1): 3–5. doi:10.1177/1745691610393980.

Bürger, R. 2000. *The Mathematical Theory of Selection, Recombination, and Mutation*. John Wiley.

Cadsby, C B, and E Maynes. 1998. "Gender and Free Riding in a Threshold Public Goods Game: Experimental Evidence." *Journal of Economic Behavior & Organization*.

Camerer, C F, and E Fehr. 2006. "When Does 'Economic Man' Dominate Social Behavior?." *Science* 311: 47–52.

Camerer, Colin. 2003. *Behavioral Game Theory: Experiments in Strategic Interaction.* Princeton University Press.

Cardenas, J C. 2007. "Wealth Inequality and Overexploitation of the Commons: Field Experiments in Colombia1." *Inequality*.

Cardenas, J C, J Stranlund, and C Willis. 2002. "Economic Inequality and Burden-Sharing in the Provision of Local Environmental Quality." *Ecological Economics* 40 (3): 379–95. doi:10.1016/S0921-8009(01)00285-3.

Cardenas, Juan-Camilo. 2003. "Real Wealth and Experimental Cooperation: Experiments in the Field Lab." *Journal of Development Economics* 70 (2): 263–89. doi:10.1016/S0304-3878(02)00098-6.

Carpenter, Jeffrey P. 2007. "Punishing Free-Riders: How Group Size Affects Mutual Monitoring and the Provision of Public Goods." *Games and Economic Behavior* 60 (1): 31–51. doi:10.1016/j.geb.2006.08.011.

Carpenter, Jeffrey, and Peter Hans Matthews. 2009. "What Norms Trigger Punishment?." *Experimental Economics* 12 (3): 272–88. doi:10.1007/s10683-009-9214-z.

Chan, K S, S Mestelman, R Moir, and R A Muller. 1996. "The Voluntary Provision of Public Goods Under Varying Income Distributions." *Canadian Journal of Economics* 29 (1): 54.

Chan, Kenneth S, Stuart Mestelman, Robert Moir, and R Andrew Muller. 1999. "Heterogeneity and the Voluntary Provision of Public Goods." *Experimental Economics* 2 (1). Kluwer Academic Publishers: 5–30. doi:10.1023/A:1009984414401.

Charité, Jimmy, Raymond Fisman, and Ilyana Kuziemko. 2015. "Reference Points and Demand for Redistribution: Experimental Evidence." *Working Paper*.

Charness, G, and M Rabin. 2002. "Understanding Social Preferences with Simple Tests." *Quarterly Journal of Economics*. doi:10.2307/4132490.

Chatterjee, Krishnendu, Damien Zufferey, and M A Nowak. 2012. "Evolutionary Game Dynamics in Populations with Different Learners." *Journal of Theoretical Biology* 301. Elsevier: 161–73.

Cherry, Todd L, Stephan Kroll, and Jason F Shogren. 2005. "The Impact of Endowment Heterogeneity and Origin on Public Good Contributions: Evidence From the Lab." *Journal of Economic Behavior & Organization* 57 (3): 357–65. doi:10.1016/j.jebo.2003.11.010.

Christakis, Nicholas a, and James H Fowler. 2007. "The Spread of Obesity in a Large Social Network Over 32 Years." *New England Journal of Medicine* 357 (4). Mass Medical Soc: 370–79.

Christakis, Nicholas a, and James H Fowler. 2008. "The Collective Dynamics of Smoking in a Large Social Network." *New England Journal of Medicine* 358 (21). Mass Medical Soc: 2249–58.

CIA World Factbook. 2014. "Literacy at Https://Www.Cia.Gov/Library/Publications/the-World-

Factbook.”

Cinyabuguma, Matthias, Talbot Page, and Louis Putterman. 2005. “Cooperation Under the Threat of Expulsion in a Public Goods Experiment.” *Journal of Public Economics* 89 (8): 1421–35. doi:10.1016/j.jpubeco.2004.05.011.

Coase, R H. 1960. “The Problem of Social Cost.” *Journal of Law and Economics* 3: 1–44.

Congressional Budget Office. 2007. “Average Federal Tax Rates.” http://www.cbo.gov/publication/42870.

Crockett, Molly J, Luke Clark, Matthew D Lieberman, Golnaz Tabibnia, and Trevor W Robbins. 2010. “Impulsive Choice and Altruistic Punishment Are Correlated and Increase in Tandem with Serotonin Depletion..” *Emotion* 10 (6). American Psychological Association: 855–62. doi:10.1037/a0019861.

Croson, RTA. 1996. “Partners and Strangers Revisited.” *Economics Letters* 53 (1): 25–32. doi:10.1016/S0165-1765(97)82136-2.

Crump, Matthew J C, John V McDonnell, and Todd M Gureckis. 2013. “Evaluating Amazon's Mechanical Turk as a Tool for Experimental Behavioral Research.” Edited by Sam Gilbert. *PLoS ONE* 8 (3): e57410. doi:10.1371/journal.pone.0057410.

Cullen, Julie Berry, Brian A Jacob, and Steven D Levitt. 2005. “The Impact of School Choice on Student Outcomes: an Analysis of the Chicago Public Schools.” *Journal of Public Economics* 89 (5). Elsevier: 729–60.

Dal Bo, Pedro. 2005. “Cooperation Under the Shadow of the Future: Experimental Evidence From Infinitely Repeated Games.” *American Economic Review* 95 (5): 1591–1604. doi:10.2307/4132766.

Davidai, S, and T Gilovich. 2015. “Building a More Mobile America—One Income Quintile at a Time.” *Perspectives on Psychological Science* 10 (1): 60–71. doi:10.1177/1745691614562005.

Dayton-Johnson, Jeff, and Pranab Bardhan. 2002. “Inequality and Conservation on the Local Commons: a Theoretical Exercise.” *The Economic Journal* 112 (481): 577–602.

Deacon, R, and P Shapiro. 1975. “Private Preference for Collective Goods Revealed Through Voting on Referenda.” *The American Economic Review*.

Deng, Lili, Wansheng Tang, and Jianxiong Zhang. 2011. “The Coevolutionary Ultimatum Game on Different Network Topologies.” *Physica a: Statistical Mechanics and Its Applications* 390 (23). Elsevier: 4227–35.

Dieckmann, Ulf, and Aacute d aacute m Kun. 2013. “Resource Heterogeneity Can Facilitate Cooperation.” *Nature Communications* 4 (September). Nature Publishing Group: 1–8. doi:10.1038/ncomms3453.

Droz, Michel, Janusz Szwabiński, and György Szabó. 2009. “Motion of Influential Players Can Support Cooperation in Prisoner’s Dilemma.” *The European Physical Journal B* 71 (4). Springer: 579–85.

Dubreuil, B. 2008. “Strong Reciprocity and the Emergence of Large-Scale Societies.” *Philosophy of the Social Sciences* 38 (2): 192–210. doi:10.1177/0048393108315509.

Durante, Ruben, Louis Putterman, and Joël van der Weele. 2014. “Preferences for Redistribution and Perception of Fairness: an Experimental Study.” *Journal of the European Economic Association* 12 (4): 1059–86. doi:10.1111/jeea.12082.

Durrett, Richard, and Simon Levin. 1994. “The Importance of Being Discrete (and Spatial).” *Theoretical Population Biology* 46 (3). Elsevier: 363–94.

Eldon, Bjarki, and John Wakeley. 2006. “Coalescent Processes When the Distribution of Offspring Number Among Individuals Is Highly Skewed.” *Genetics* 172 (4). Genetics Soc America: 2621–33.

Ertan, Arhan, Page, Talbot, and Louis Putterman. 2009. “Who to Punish? Individual Decisions and Majority Rule in Mitigating the Free Rider Problem.” *European Economic Review* 53 (5): 495–511. doi:10.1016/j.euroecorev.2008.09.007.

Essock-Vitale, Susan M. 1984. “The Reproductive Success of Wealthy Americans.” *Ethology and Sociobiology* 5 (1). Elsevier: 45–49.

Esteban, Joan, and Debraj Ray. 2001. “Collective Action and the Group Size Paradox.” *American Political Science Association* 95 (03). Cambridge University Press: 663–72. doi:10.1017/S0003055401003124.

Fehr, E, and S Gächter. 2002. “Altruistic Punishment in Humans.” *Nature* 415 (6868): 137–40. doi:10.1038/415137a.

Fehr, Ernst, and Simon Gächter. 2000. “Cooperation and Punishment in Public Goods Experiments.” *American Economic Review* 90 (4): 980–94.

Fischbacher, Urs, Simon Gächter, and Ernst Fehr. 2001. “Are People Conditionally Cooperative? Evidence From a Public Goods Experiment.” *Economics Letters* 71 (3): 397–404. doi:10.1016/S0165-1765(01)00394-9.

Fisher, Joseph, R Mark Isaac, Jeffrey W Schatzberg, and James M Walker. 1995. “Heterogenous Demand for Public Goods: Behavior in the Voluntary Contributions Mechanism.” *Public Choice* 85 (3-4). Springer: 249–66.

Forsythe, R, J L Horowitz, N E Savin, and M Sefton. 1994. “Fairness in Simple Bargaining Experiments.” *Games and Economic …*.

Fosgaard, Toke Reinholt, Las Garn Hansen, and Erik Wengström. 2011. “Framing and Misperceptions in a Public Good Experiment.” *Working Paper*, 1–35.

Fowler, J H, and N A Christakis. 2010. “Cooperative Behavior Cascades in Human Social Networks.” *Proceedings of the National Academy of Sciences* 107 (12): 5334–38. doi:10.1073/pnas.0913149107.

Frean, Marcus, Paul B Rainey, and Arne Traulsen. 2013. “The Effect of Population Structure on the Rate of Evolution.” *Proceedings of the Royal Society B: Biological Sciences* 280 (1762). The Royal Society: 20130211.

Fu, Feng, C Hauert, M A Nowak, and Long Wang. 2008. “Reputation-Based Partner Choice Promotes Cooperation in Social Networks.” *Physical Review E* 78 (2). APS: 026117.

Fu, Feng, L-H Liu, and Long Wang. 2007. “Evolutionary Prisoner's Dilemma on Heterogeneous Newman-Watts Small-World Network.” *The European Physical Journal B* 56 (4). Springer: 367–72.

Fudenberg, D, and Lorens A Imhof. 2006. “Imitation Processes with Small Mutations.” *Journal of Economic Theory* 131 (1). Elsevier: 251–62.

Fudenberg, D, David G Rand, and Anna Dreber. 2012. “Slow to Anger and Fast to Forgive: Cooperation in an Uncertain World.” *American Economic Review* 102 (2): 720–49. doi:10.1257/aer.102.2.720.

Gächter, S, E Renner, and M Sefton. 2008. “The Long-Run Benefits of Punishment.” *Science* 322 (5907): 1510–10. doi:10.1126/science.1164744.

Gächter, Simon, Friederike Mengel, Elias Tsakas, and Alexander Vostroknutov. 2014. “Growth and Inequality in Public Good Games.” *Working Paper*, July. doi:10.2139/ssrn.2351717.

Grechenig, K, and A Nicklisch. 2010. “Punishment Despite Reasonable Doubt—a Public Goods Experiment with Sanctions Under Uncertainty.” *Journal of Empirical Legal ….*

Grinstead, Charles Miller, and James Laurie Snell. 2012. *Introduction to Probability*. American Mathematical Society.

Grujić, Jelena, Burcu Eke, Antonio Cabrales, José A Cuesta, and Angel Sánchez. 2012. “Three Is a Crowd in Iterated Prisoner's Dilemmas: Experimental Evidence on Reciprocal Behavior.” *Scientific Reports* 2 (September): 1–7. doi:10.1038/srep00638.

Hackett, S, E Schlager, and J Walker. 1994. “The Role of Communication in Resolving Commons Dilemmas: Experimental Evidence with Heterogeneous Appropriators.” *Journal of Environmental Economics and ….*

Hardin, Garrett. 1968. “The Tragedy of the Commons.” *Science* 162 (3859). American Association for the Advancement of Science: 1243–48. doi:10.1126/science.162.3859.1243.

Hauser, Oliver P, Arne Traulsen, and M A Nowak. 2014. “Heterogeneity in Background Fitness Acts as a Suppressor of Selection.” *Journal of Theoretical Biology* 343 (February): 178–85. doi:10.1016/j.jtbi.2013.10.013.

Haynes, Laura, Owain Service, Ben Goldacre, and David Torgerson. 2012. "Test, Learn, Adapt: Developing Public Policy with Randomised Controlled Trials." *Working Paper*.

Helbing, D. 2010. *Quantitative Sociodynamics: Stochastic Methods and Models of Social Interaction Processes*. Springer.

Hendriks, Achim. 2012. "SoPHIE - Software Platform for Human Interaction Experiments." *Working Paper*.

Henrich, Joseph, Jean Ensminger, Richard McElreath, Abigail Barr, Clark Barrett, Alexander Bolyanatz, Juan-Camilo Cardenas, et al. 2010. "Markets, Religion, Community Size, and the Evolution of Fairness and Punishment." *Science* 327 (5972). American Association for the Advancement of Science: 1480–84. doi:10.1126/science.1182238.

Herrmann, B, C Thöni, and S Gächter. 2008. "Antisocial Punishment Across Societies." *Science* 319 (5868): 1362–67. doi:10.1126/science.1153808.

Hill, Alison L, David G Rand, M A Nowak, and Nicholas a Christakis. 2010. "Emotions as Infectious Diseases in a Large Social Network: the SISa Model." *Proceedings of the Royal Society B: Biological Sciences* 277 (1701). The Royal Society: 3827–35.

Hofbauer, Josef, and K Sigmund. 2003. "Evolutionary Game Dynamics." *Bulletin of the American Mathematical Society* 40 (4): 479–519.

Hofmeyr, A, J Burns, and M Visser. 2007. "Income Inequality, Reciprocity and Public Good Provision: an Experimental Analysis." *South African Journal of ...*.

Holcombe, R G. 1989. "The Median Voter Model in Public Choice Theory." *Public Choice*.

Horton, John J, David G Rand, and Richard J Zeckhauser. 2011. "The Online Laboratory: Conducting Experiments in a Real Labor Market." *Experimental Economics* 14 (3): 399–425. doi:10.1007/s10683-011-9273-9.

Imhof, Lorens A, and M A Nowak. 2006. "Evolutionary Game Dynamics in a Wright-Fisher Process." *Journal of Mathematical Biology* 52 (5). Springer: 667–81.

Isaac, R Mark, and James M Walker. 1988. "Communication and Free-Riding Behavior: the Voluntary Contribution Mechanism." *Economic Inquiry* 26 (4).

Jacquet, Jennifer. 2015. *Is Shame Necessary? New Uses for an Old Tool*. Pantheon.

Jacquet, Jennifer, Kristin Hagel, C Hauert, Jochem Marotzke, Torsten Röhl, and Manfred Milinski. 2013. "Intra- and Intergenerational Discounting in the Climate Game." *Nature Climate Change* 3 (12). Nature Publishing Group: 1025–28. doi:10.1038/nclimate2024.

Janssen, Marco A, Robert Holahan, Allen Lee, and Elinor Ostrom. 2010. "Lab Experiments for the Study of Social-Ecological Systems." *Science* 328 (5978): 613–17.

Kahneman, D. 2003. “Maps of Bounded Rationality: Psychology for Behavioral Economics.” *American Economic Review*.

Kamei, Kenju, Louis Putterman, and Jean-robert Tyran. 2014. “State or Nature? Endogenous Formal Versus Informal Sanctions in the Voluntary Provision of Public Goods.” *Experimental Economics*, June. doi:10.1007/s10683-014-9405-0.

Keister, L A, and S Moller. 2000. “Wealth Inequality in the United States.” *Annual Review of Sociology*.

Keser, Claudia, Andreas Markstädter, and Martin Schmidt. 2014. “Mandatory Minimum Contributions, Heterogenous Endowments and Voluntary Public-Good Provision.” *Working Paper*, December. doi:10.2139/ssrn.2539601.

Keser, Claudia, Andreas Markstädter, Martin Schmidt, and Cornelius Schnitzler. 2014. “Social Cost of Inequality - Heterogeneous Endowments in Public-Good Experiments.” *Working Paper*, October, 1–28.

Kiatpongsan, S, and M I Norton. 2014. “How Much (More) Should CEOs Make? a Universal Desire for More Equal Pay.” *Perspectives on Psychological Science* 9 (6). SAGE Publications: 587–93. doi:10.1177/1745691614549773.

Kokko, Hanna. 1999. “Competition for Early Arrival in Migratory Birds.” *Journal of Animal Ecology* 68 (5). Wiley Online Library: 940–50.

Kraft-Todd, Gordon, Erez Yoeli, Syon Bhanot, and David Rand. 2015. “Promoting Cooperation in the Field.” *Current Opinion in Behavioral Sciences* 3 (June). Elsevier Ltd: 96–101. doi:10.1016/j.cobeha.2015.02.006.

Kuziemko, Ilyana, Michael I Norton, Emmanuel Saez, and Stefanie Stantcheva. 2015. “How Elastic Are Preferences for Redistribution? Evidence From Randomized Survey Experiments.” *American Economic Review* 105 (4). Cambridge, MA: National Bureau of Economic Research: 1478–1508. http://www.nber.org/papers/w18865.

Lago-Peñas, Ignacio, and Santiago Lago-Peñas. 2010. “The Determinants of Tax Morale in Comparative Perspective: Evidence From European Countries.” *European Journal of Political Economy* 26 (4). Elsevier B.V.: 441–53. doi:10.1016/j.ejpoleco.2010.06.003.

Lessard, Sabin. 2007. “Cooperation Is Less Likely to Evolve in a Finite Population with a Highly Skewed Distribution of Family Size.” *Proceedings of the Royal Society B: Biological Sciences* 274 (1620). The Royal Society: 1861–65.

Levin, Simon. 2007. *Fragile Dominion*. Basic Books.

Levin, Simon A. 2009. *Games, Groups, and the Global Good*. Edited by Simon A Levin. Berlin, Heidelberg: Springer. doi:10.1007/978-3-540-85436-4.

Lieberman, E, C Hauert, and M A Nowak. 2005. “Evolutionary Dynamics on Graph.” *Nature*

433 (7023): 312–16.

Mason, Winter, and Siddharth Suri. 2011. “Conducting Behavioral Research on Amazon’s Mechanical Turk.” *Behavior Research Methods* 44 (1): 1–23. doi:10.3758/s13428-011-0124-6.

Masuda, Naoki, and H Ohtsuki. 2009. “Evolutionary Dynamics and Fixation Probabilities in Directed Networks.” *New Journal of Physics* 11 (3). IOP Publishing: 033012.

McDonald, John H. 2014. *Handbook of Biological Statistics*. Sparky House Publishing.

McNamara, J M, Z Barta, and A I Houston. 2004. “Variation in Behaviour Promotes Cooperation in the Prisoner's Dilemma Game.” *Nature*.

Milinski, M, and D Semmann. 2006. “Stabilizing the Earth's Climate Is Not a Losing Game: Supporting Evidence From Public Goods Experiments.” In.

Milinski, M, and R D Sommerfeld. 2008. “The Collective-Risk Social Dilemma and the Prevention of Simulated Dangerous Climate Change.” In.

Milinski, M, D Semmann, T C M Bakker, and H J Krambeck. 2001. “Cooperation Through Indirect Reciprocity: Image Scoring or Standing Strategy?.” *Proceedings of the Royal Society B: Biological Sciences* 268 (1484): 2495–2501. doi:10.1098/rspb.2001.1809.

Milinski, Manfred, Dirk Semmann, and Hans-Jürgen Krambeck. 2002. “Reputation Helps Solve the ‘Tragedy of the Commons’.” *Nature* 415 (6870). Nature Publishing Group: 424–26. doi:10.1038/415424a.

Milinski, Manfred, Torsten Röhl, and Jochem Marotzke. 2011. “Cooperative Interaction of Rich and Poor Can Be Catalyzed by Intermediate Climate Targets.” *Climatic Change* 109 (3-4): 807–14. doi:10.1007/s10584-011-0319-y.

Misenhelter, Michael D, and John T Rotenberry. 2000. “Choices and Consequences of Habitat Occupancy and Nest Site Selection in Sage Sparrows.” *Ecology* 81 (10). Eco Soc America: 2892–2901.

Moran, Patrick Alfred Pierce. 1962. “The Statistical Processes of Evolutionary Theory..” *The Statistical Processes of Evolutionary Theory.* Clarendon Press; Oxford University Press.

Mueller, D C. 1979. *Public Choice*. Cambridge University Press.

Mulder, M B, S Bowles, T Hertz, A Bell, J Beise, G Clark, I Fazzio, et al. 2009. “Intergenerational Wealth Transmission and the Dynamics of Inequality in Small-Scale Societies.” *Science* 326 (5953): 682–88. doi:10.1126/science.1178336.

Nikiforakis, Nikos, and Hans-Theo Normann. 2007. “A Comparative Statics Analysis of Punishment in Public-Good Experiments.” *Experimental Economics* 11 (4): 358–69. doi:10.1007/s10683-007-9171-3.

Nishi, Akihiro, Hirokazu Shirado, David G Rand, and Nicholas a Christakis. 2015. "Inequality and Visibility of Wealth in Experimental Social Networks." *Nature*, September, 1–14. doi:10.1038/nature15392.

Norton, Michael I, and Dan Ariely. 2011. "Building a Better America—One Wealth Quintile at a Time." *Perspectives on Psychological Science* 6 (1). SAGE Publications: 9–12. doi:10.1177/1745691610393524.

Norwegian Tax Administration. 2015. "Norway Income Database." Skatteetaten. http://www.skatteetaten.no/nn/Person/Skatteoppgjer/Sok-i-skattelistene/Hva-star-i-skattelistene/.

Nowak, M A. 2006a. *Evolutionary Dynamics*. Harvard University Press.

Nowak, M A. 2006b. "Five Rules for the Evolution of Cooperation." *Science* 314 (5805): 1560–63. doi:10.1126/science.1133755.

Nowak, M A, A Sasaki, C Taylor, and D Fudenberg. 2004. "Emergence of Cooperation and Evolutionary Stability in Finite Populations." *Nature* 428 (6983): 646–50.

Nowak, M A, and K Sigmund. 1992. "Tit for Tat in Heterogeneous Populations." *Nature*.

Nowak, M A, and K Sigmund. 2004. "Evolutionary Dynamics of Biological Games." *Science* 303 (5659): 793–99. doi:10.1126/science.1093411.

Nowak, M A, Corina E Tarnita, and Edward O Wilson. 2010. "The Evolution of Eusociality." *Nature* 466 (7310). Nature Publishing Group: 1057–62. doi:10.1038/nature09205.

Nowak, M A, F Michor, and Y Iwasa. 2003. "The Linear Process of Somatic Evolution." In.

Ogburn, William F, and Delvin Peterson. 1916. "Political Thought of Social Classes." *Political Science Quarterly*. JSTOR, 300–317.

Ohtsuki, H, C Hauert, E Lieberman, and M A Nowak. 2006. "A Simple Rule for the Evolution of Cooperation on Graphs and Social Networks." *Nature* 441 (7092): 502–5. doi:10.1038/nature04605.

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

Ostrom, E, J Walker, and R Gardner. 1992. "Covenants with and Without a Sword: Self-Governance Is Possible.." *American Political Science ….*

Ostrom, Elinor. 1990. *Governing the Commons: the Evolution of Institutions for Collective Action*. Cambridge University Press.

Oullier, Olivier. 2013. "Behavioural Insights Are Vital to Policy-Making." *Nature* 501: 463.

Pacheco, Jorge M, Francisco C Santos, Max O Souza, and Brian Skyrms. 2011. "Evolutionary

Dynamics of Collective Action.” In *The Mathematics of Darwins Legacy*, 119–38. Springer.

Paolacci, G, and J Chandler. 2014. “Inside the Turk: Understanding Mechanical Turk as a Participant Pool.” *Current Directions in Psychological Science* 23 (3): 184–88. doi:10.1177/0963721414531598.

Perc, Matjaz, and Attila Szolnoki. 2008. “Social Diversity and Promotion of Cooperation in the Spatial Prisoner’s Dilemma Game.” *Physical Review E* 77 (1). APS: 011904.

Perc, Matjaz, and Attila Szolnoki. 2010. “Coevolutionary Games—a Mini Review.” *BioSystems* 99 (2). Elsevier: 109–25.

Perez-Truglia, Ricardo, and Ugo Troiano. 2015. “Shaming Tax Delinquents: Theory and Evidence From a Field Experiment in the United States.” *Working Paper*, September, 1–64.

Poncela, Julia, Jesús Gómez-Gardeñes, Arne Traulsen, and Yamir Moreno. 2009. “Evolutionary Game Dynamics in a Growing Structured Population.” *New Journal of Physics* 11 (8). IOP Publishing: 083031.

Putterman, Louis, Jean-robert Tyran, and Kenju Kamei. 2011. “Public Goods and Voting on Formal Sanction Schemes.” *Journal of Public Economics* 95 (9-10): 1213–22. doi:10.1016/j.jpubeco.2011.05.001.

Quandl.com. 2014. “Quandl Data Browser at Http://Www.Quandl.com.”

Rand, David G, and M A Nowak. 2013. “Human Cooperation.” *Trends in Cognitive Sciences* 17 (8). Elsevier Ltd: 413–25. doi:10.1016/j.tics.2013.06.003.

Rand, David G, Anna Dreber, Tore Ellingsen, Drew Fudenberg, and M A Nowak. 2009. “Positive Interactions Promote Public Cooperation.” *Science* 325: 1272–75.

Rand, David G, C E Tarnita, and H Ohtsuki. 2013. “Evolution of Fairness in the One-Shot Anonymous Ultimatum Game.” In. doi:10.1073/pnas.1214167110/-/DCSupplemental.

Rand, David G, Joshua D Greene, and M A Nowak. 2012. “Spontaneous Giving and Calculated Greed.” *Nature* 489 (7416): 427–30. doi:10.1038/nature11467.

Rand, David G, M A Nowak, James H Fowler, and Nicholas a Christakis. 2014. “Static Network Structure Can Stabilize Human Cooperation.” *Proceedings of the National Academy of Sciences* 111 (48): 17093–98. doi:10.1073/pnas.1400406111.

Rand, David G, Samuel Arbesman, and Nicholas a Christakis. 2011. “Dynamic Social Networks Promote Cooperation in Experiments with Humans.” *Proceedings of the National Academy of Sciences* 108 (48). National Acad Sciences: 19193–98. doi:10.1073/pnas.1108243108.

Reuben, E, and A Riedl. 2008. “Public Goods Provision and Sanctioning in Privileged Groups.” *Journal of Conflict Resolution* 53 (1). SAGE Publications: 72–93. doi:10.1177/0022002708322361.

Reuben, Ernesto, and Arno Riedl. 2013. “Enforcement of Contribution Norms in Public Good Games with Heterogeneous Populations.” *Games and Economic Behavior* 77 (1). Elsevier Inc.: 122–37. doi:10.1016/j.geb.2012.10.001.

Rockenbach, Bettina, and Manfred Milinski. 2006. “The Efficient Interaction of Indirect Reciprocity and Costly Punishment.” *Nature* 444 (7120): 718–23. doi:10.1038/nature05229.

Santos, Francisco C, and Jorge M Pacheco. 2005. “Scale-Free Networks Provide a Unifying Framework for the Emergence of Cooperation.” *Physical Review Letters* 95 (9). APS: 098104.

Santos, Francisco C, Flavio L Pinheiro, Tom Lenaerts, and Jorge M Pacheco. 2012. “The Role of Diversity in the Evolution of Cooperation.” *Journal of Theoretical Biology* 299. Elsevier: 88–96.

Santos, Francisco C, Jorge M Pacheco, and Tom Lenaerts. 2006. “Evolutionary Dynamics of Social Dilemmas in Structured Heterogeneous Populations.” *Proceedings of the National Academy of Sciences* 103 (9). National Acad Sciences: 3490–94.

Santos, Francisco C, Marta D Santos, and Jorge M Pacheco. 2008. “Social Diversity Promotes the Emergence of Cooperation in Public Goods Games.” *Nature* 454 (7201): 213–16. doi:10.1038/nature06940.

Sefton, Martin, Robert Shupp, and James M Walker. 2007. “The Effect of Rewards and Sanctions in Provision of Public Goods.” *Economic Inquiry* 45 (4). Blackwell Publishing Inc: 671–90. doi:10.1111/j.1465-7295.2007.00051.x.

Smith, John Maynard. 1993. “The Theory of Evolution.” Cambridge University Press.

Sutter, Matthias, Stefan Haigner, and Martin G Kocher. 2010. “Choosing the Carrot or the Stick? Endogenous Institutional Choice in Social Dilemma Situations.” *Review of Economic Studies* 77 (4): 1540–66. doi:10.1111/j.1467-937X.2010.00608.x.

Szabó, György, and Gabor Fath. 2007. “Evolutionary Games on Graphs.” *Physics Reports* 446 (4). Elsevier: 97–216.

Tarnita, Corina E, Clifford H Taubes, and M A Nowak. 2013. “Evolutionary Construction by Staying Together and Coming Together.” *Journal of Theoretical Biology* 320. Elsevier: 10–22.

Tarnita, Corina E, Nicholas Wage, and M A Nowak. 2011. “Multiple Strategies in Structured Populations.” *Proceedings of the National Academy of Sciences* 108 (6). National Acad Sciences: 2334–37.

Tarnita, Corina E, Tibor Antal, H Ohtsuki, and M A Nowak. 2009. “Evolutionary Dynamics in Set Structured Populations.” *Proceedings of the National Academy of Sciences* 106 (21). National Acad Sciences: 8601–4.

Tavoni, A, A Dannenberg, and G Kallis. 2011. “Inequality, Communication, and the Avoidance of Disastrous Climate Change in a Public Goods Game.” In. doi:10.1073/pnas.1102493108/-/DCSupplemental.

Taylor, C, and M A Nowak. 2006. “Evolutionary Game Dynamics with Non-Uniform Interaction Rates.” *Theoretical Population Biology* 69 (3). Elsevier: 243–52.

The Economist Intelligence Unit. 2012. “Democracy Index at Http://Www.Eiu.com/Public/Topical_Report.Aspx?Campaignid=DemocracyIndex12.”

Traulsen, A, and M A Nowak. 2006. “Evolution of Cooperation by Multilevel Selection.” In.

Traulsen, Arne, Jens Christian Claussen, and C Hauert. 2005. “Coevolutionary Dynamics: From Finite to Infinite Populations.” *Physical Review Letters* 95 (23). APS: 238701.

Traulsen, Arne, Jorge M Pacheco, and M A Nowak. 2007. “Pairwise Comparison and Selection Temperature in Evolutionary Game Dynamics.” *Journal of Theoretical Biology* 246 (3). Elsevier: 522–29.

Traulsen, Arne, M A Nowak, and Jorge M Pacheco. 2007. “Stochastic Payoff Evaluation Increases the Temperature of Selection.” *Journal of Theoretical Biology* 244 (2). Elsevier: 349–56.

U.S. Census Bureau. 2013. “Income and Poverty in the United States.” http://www.census.gov/hhes/www/income/data/incpovhlth/2013/table4.pdf.

Ule, A, A Schram, A Riedl, and T N Cason. 2009. “Indirect Punishment and Generosity Toward Strangers.” *Science* 326 (5960): 1701–4. doi:10.1126/science.1178883.

Van der Heijden, ECM, and JHM Nelissen. 1998. “Transfers and the Effect of Monitoring in an Overlapping-Generations Experiment.” *European Economic …*.

Van Dijk, E, and H Wilke. 1994. “Asymmetry of Wealth and Public Good Provision.” *Social Psychology Quarterly*.

Varughese, George, and Elinor Ostrom. 2001. “The Contested Role of Heterogeneity in Collective Action: Some Evidence From Community Forestry in Nepal.” *World Development* 29 (5). Elsevier: 747–65.

Vasconcelos, Vítor V, Francisco C Santos, and Jorge M Pacheco. 2013. “A Bottom-Up Institutional Approach to Cooperative Governance of Risky Commons.” *Nature Climate Change* 3 (0). Nature Publishing Group: 1–5. doi:10.1038/nclimate1927.

Voelkl, Bernhard. 2010. “The ‘Hawk-Dove’Game and the Speed of the Evolutionary Process in Small Heterogeneous Populations.” *Games* 1 (2). Molecular Diversity Preservation International: 103–16.

Wade-Benzoni, K A. 2002. “A Golden Rule Over Time: Reciprocity in Intergenerational

Allocation Decisions." *Academy of Management Journal*.

Wade-Benzoni, K A, and L P Tost. 2009. "The Egoism and Altruism of Intergenerational Behavior." *Personality and Social Psychology Review* 13 (3): 165–93.

Wakeley, John. 2008. "Coalescent Theory: an Introduction." Roberts & Company Publishers Greenwood Village, Colorado.

Walker, J M, R Gardner, and A Herr. 2000. "Collective Choice in the Commons: Experimental Results on Proposed Allocation Rules and Votes." *The Economic Journal* 110 (460): 212–34.

Wang, Jing, Feng Fu, and Long Wang. 2010. "Effects of Heterogeneous Wealth Distribution on Public Cooperation with Collective Risk." *Physical Review E* 82 (1). APS: 016102.

Weber, Elke U, and Eric J Johnson. 2012. "Psychology and Behavioral Economics Lessons for the Design of a Green Growth Strategy." *Working Paper*, October, 1–50.

Wedekind, Claus, and Manfred Milinski. 2000. "Cooperation Through Image Scoring in Humans." *Science* 288 (5467): 850–52. doi:10.1126/science.288.5467.850.

Weibull, Jörgen W. 1997. "Evolutionary Game Theory." MIT press.

Williamson, O E. 1985. *The Economic Intstitutions of Capitalism*. Simon and Schuster.

Wolff, Edward N. 2002. "Inheritances and Wealth Inequality, 1989-1998." *American Economic Review*. JSTOR, 260–64.

World Bank. 2004. "WBI Governance & Anti-Corruption"

World Bank. 2013. "GDP Ranking at Http://Data.Worldbank.org/Data-Catalog/GDP- Ranking-Table."

World Bank. "Gini Index." http://wdi.worldbank.org/table/2.9.

World Health Organization. 2013. "Life Expectancy at Http://Www.Who.Int/Gho/Mortality_Burden_Disease/Life_Tables/Situation_Trends/en."

WorldEnergy.org. "Energy Sustainability Index at www.Worldenergy.org/Data/Sustainability-Index."

Yoeli, Erez, Moshe Hoffman, David G Rand, and M A Nowak. 2013. "Powering Up with Indirect Reciprocity in a Large-Scale Field Experiment." *Proceedings of the National Academy of Sciences* 110: 10424–29. doi:10.1073/pnas.1301210110.