# Essays on Innovation and Mutual Insurance

Julia Katharina Wirtz

A thesis presented for the degree of
Doctor of Philosophy

Department of Economics
University of Warwick

May 2014

# Contents

# List of Figures

## Acknowledgements

I would like to to express my heartfelt gratitude to my supervisors Motty Perry and Andrés Carvajal for their continuous support and guidance. Both Andrés and Motty are the kindest and most patient supervisors anyone could hope for.

Parts of this dissertation have been presented at the PET meeting in Lisbon, at the 7th Workshop on Economic Theory at the University of Vigo, at the University of Norwich and at the University of Warwick where I received many useful comments. I am also indebted for insightful discussions to Marina Halac, Jacob Glazer, Omer Moav, Ilan Kremer, Dan Bernhardt, and Phil Reny. All remaining errors and omissions are my responsibility.

Financial Support from the ESRC and from the Department of Economics at Warwick is gratefully acknowledged. Beyond the financial support, I am also grateful to the department for providing a both intellectually stimulating and sociable environment.

I am grateful to my family for their moral and material support during my extensive studies and to my friends at Warwick for spirited discussions and for making my time here so enjoyable. Finally, would like to thank Alessandro Iaria for continuous encouragement, patience and inspiration.

## Declaration and Inclusion of Material from a Prior Thesis

This thesis is submitted to the University of Warwick in support of my application for the degree of Doctor of Philosophy. It has been composed by myself and has not been submitted in any previous application for any degree. The work presented was carried out by the author except in the cases outlined below:

Chapters 2 and 3 were written jointly with Alessandro Iaria. The basic idea and all main intuitions and results of chapter 2 were developed in joint discussions while Alessandro was mainly responsible for the final editing of the associated proofs. The basic idea of chapter 3 was similarly developed in joint discussions, while I obtained all results and was responsible for the final editing.

## Introduction

This thesis is composed of three chapters. While chapter 1 stands on it's own, chapters 2 and 3 are related in topic and grew out of two parts of a single paper. Hence, even though they constitute largely independent treatments of separate questions they have a unique introduction and conclusion.

Chapter 1 considers a dynamic tournament setting where feedback gives agents the possibility to learn about the productivity of a technology they are using. If the technology proves unsatisfactory, they have the possibility of switching to a different one. However, it is shown that full feedback does not lead to the efficient technology choice in a tournament setting. Risk neutral agents will behave as risk averse in some cases and risk loving in others. It is shown that the inefficiency can be ameliorated by giving the later period more weight for the allocation of the tournament prize. In a setting with effort. feedback will induce the agents to exert higher effort in the presence of learning. Finally, the efficient technology choice can be achieved using partial feedback.

Chapters 2 and 3 investigate the phenomenon of kindness towards strangers. Seemingly altruistic behaviour can follow from purely selfish motives if agents face risk. In a repeated dictator game setting, a charitable equilibrium can be sustained if dictators have a positive probability to change roles, even with anonymous transactions. This also holds if behaviour cannot be monitored perfectly. The main driving factor for charitable behaviour is the desire to sustain the social norm of kindness, from which the charitable agent herself might benefit in the future. This can be interpreted as an informal insurance arrangement in the absence of enforceable contracts. Furthermore, we examine how cooperation is complicated by inequality, in terms of heterogeneous risk exposure. We study how more persistent differences in risk make cooperation increasingly hard to achieve. Moreover, heterogeneity can lead to a fragmentation of society, where cooperation is only possible within subgroups, leading to losses in welfare. The model allows for interesting interpretations of social divisions in societies of varying heterogeneity.

# Chapter 1. Feedback and Learning in Tournaments

## 1. Introduction

This paper studies feedback in dynamic tournaments and the resulting behaviour of the competitors. Tournaments are competitions were the rewards of competitors depends on their ranking in relation to each other rather than solely on individual performance. Tournaments are pervasive in economic and social settings. Employees compete for a promotion, assistant professors compete for a limited number of tenured positions, students compete to be on top of their class, athletes compete for medals and research teams compete for patents in R&D races.

This paper studies the optimal design of dynamic tournaments that last over more than one period. The studies settings where the competitors have a choice between different strategies, projects or technologies. They can learn about their quality and possibly switch to a different strategy during the tournament. While the literature on tournaments has largely focused on the amount of effort agents provide in a tournament, i.e. *how hard* they work, this paper focuses on learning and strategy choice of agents, i.e. *how* they work.

There are various reasons for the existence of tournaments. In the presence of large correlated shocks tournaments can be an optimal incentive mechanism. In other cases, tournaments are the only possible mechanism. This can be the case if there is imperfect monitoring, such that only the rank order of agents is observable but not a measure of individual performance. If the performance of the agents is not verifiable, e.g. when it is subject to the subjective judgement of a principal, tournaments may be the only credible incentive mechanism. Furthermore there are situations where tournaments are not optimal but externally imposed. This can be the case if there is only a limited number of indivisible prizes as with promotions and sometimes tenured positions. Another example is when the competitors genuinely care about the rank order itself as with student rankings or sports competitions.

The earliest theoretical contributions considered mostly static tournaments. Lazear and Rosen (1981), Green and Stokey (1983) as well as Nalebuff and Stiglitz (1983) compare the efficiency of tournaments and individual incentive contracts. They find that tournaments can be superior for risk-averse agents in the presence of unobserved correlated shocks.

More recently, a literature considering dynamic tournaments has emerged. If agents exert effort over time, the principal has to decide various aspects in the design of the tournament. First, there is the decision whether to monitor interim performance. Aoyagi (2010); Ederer (2010) and Goltsman and Mukherjee (2011) study whether the principal should then give feedback during the tournaments and reveal the interim results to the agents. Aoyagi finds that either a policy of full feedback, where the principal publicly reveals output, or a policy of no feedback at all will be optimal, depending on the shape of agent's cost function. Ederer considers a setting where agents differ in ability and restricts his consideration to the alternative policies of full and no feedback. Likewise he finds that the curvature of the cost function determines which policy is superior in a setting where ability enters additively in the production function. When ability enters multiplicatively the outcome of a full feedback policy improves due to efficient sorting. Goltsman and Mukherjee find that partial feedback is optimal in a setting with a binary output. Gershkov and Perry (2009) study whether the principal should conduct a midterm review and how much weight it should carry. The feedback policy is fixed as it is assumed that the result of the midterm review is public. They find that it is always optimal to conduct a midterm review given the correct aggregation rule for midterm review and final outcome. According to the optimal aggregation rule, the second period weight should increase with the effect of first period effort on final output.

There exists a related literature on feedback in single agent settings. Lizzeri et al. (2002) finds that it is often not optimal to give feedback Given optimal incentives effort can be induced more cheaply without effort. Fuchs (2006) studies a setting where output is privately observed by the principal over multiple periods and also finds that it might be optimal not to provide any feedback. Instead the agent is fired if output falls below a threshold.

The existing literature almost exclusively focuses on efficiency of tournaments and feedback mechanisms in incentivising effort. In addition, Ederer considers that feedback allows agents to learn about their ability. However, a different function of feedback is largely ignored by the economics literature. Feedback allows agents not only to learn about their basic ability. Feedback also allows the agents to assess if a strategy, project or technology they are using is suitable or if it should be changed. This question is relevant in many settings. Students might reconsider their strategy for studying after a low mark in a midterm test, assistant professors might reconsider their research topic, investment bankers could adjust their portfolio and R&D teams might change their method. Indeed this aspect is often acknowledged, but seen as unproblematic[1] or is at least not considered in detail[2]. This paper provides a more thorough analysis of feedback and strategy choice in tournament settings. This issue is related to a growing literature on incentivising learning and innovation.[3]

While the literature on feedback and tournaments has largely assessed feedback negatively or at least ambiguously, taking into account the importance of feedback for learning and adjusting strategies leads to a more favourable view. Without feedback, learning about the the quality of the technology used is impossible. However, it is shown that there can be inefficiencies not only of effort but also of strategy choice in a tournament setting. With full feedback agents might choose inefficiently risky or inefficiently safe technologies. This could provide an explanation for investment bankers investing in overly risky projects in a competitive

[1] Lizzeri et al. (p.2) write: "To the extent that providing feedback on performance helps individuals do their jobs better, or plan their futures better, it is beneficial. But what effects does performance feedback have on incentives and motivation?"

[2] Nalebuff and Stiglitz mention that competitive compensation schemes can potentially induce risk-averse agents to choose riskier techniques, but they do not explicitly analyse this aspect in the context of tournaments.

[3] See e.g. Manso (2011) for optimal contracts with a single agent or Ederer (2013) for multiple agents and also Moscarini and Squintani (2010); Halac et al. (2012); Gomes et al. (2013) and Kremer et al. (2013).

setting or why assistant professors might stick with a known but not very fruitful research topic instead of exploring a new one. Furthermore, it is shown that inefficiencies in technology choice can be ameliorated by putting greater weight on later periods. This could provide a rationale for being more lenient with a beginner and giving the results of the first period less importance for the allocation of the prize. Lastly, if partial feedback is possible, the principal can induce the optimal technology choice by giving recommendations.

Section 2 introduces the problem of inefficient strategy choice in a basic model. Section 3 considers the optimal allocation rule for the prize. Section 4 then examines the interaction between strategy choice and effort in settings with and without effort. Finally, section 5 studies strategy choice under partial feedback.

## 2. Basic Model

In the basic model we contrast the extreme policies of full public feedback and no feedback at all. Two agents $(i = a, b)$ compete in a tournament over two periods $(t = 1, 2)$ for a fixed prize of 1. In each period the agents choose a technology from a continuum of ex ante identical technologies, with productivity $\theta \sim N(\mu, \sigma^2)$. The first period output of agent $i$ is $x_1^i = \theta_1^i + \varepsilon_1^i$, the sum of the productivity of their chosen technology and the error term $\varepsilon^i \sim N(0, 1)$. Before the second period the agents choose whether to keep the tested technology or to switch to a new one. Second period output is then simply the productivity of the used technology: $x_2^i = \theta_2^i$. The prize goes to the agent with the higher total output $x^i = x_1^i + x_2^i = \theta_1^i + \varepsilon_1^i + \theta_2^i$. The principal privately observes the output and decides whether to give the agents feedback after period 1 and publicly disclose $x_1^a, x_1^b$.[4] The principal is risk neutral and maximises aggregate total output $x^a + x^b$. The agents are risk neutral. Before the second period the agents choose an action $a^i \epsilon \{K, S\}$. They choose whether to keep (K) the tested technology $(\theta_2^i = \theta_K^i)$ or to switch (S) to a new one $(\theta_2^i = \theta_S^i)$ to maximise their expected payoff which is

[4] In order to keep the problem tractable it is assumed that the principal can commit to giving truthful feedback. This is relaxed in Section5. See Goltsman and Mukherjee (2011) for the optimal partial feedback in a setting without learning and with a binary outcome.

equal to the probability of winning

$$U^i = \Pr\left(x^i > x^j\right) = F_{\Delta x}\left(0\right),$$

where $\Delta x_1 = x_1^i - x_1^j$. If the agents do not receive feedback, they clearly cannot condition their switching decision on the first period output. For K: $\Delta x \sim N\left(0, 8\sigma^2 + 2\right)$, while for S: $\Delta x \sim N\left(0, 4\sigma^2 + 2\right)$. The probability of winning with K and S is always $\frac{1}{2}$. Hence the agents are indifferent. It is assumed that agents switch as their output has lower variance.

If the agents receive feedback, they update their beliefs about the used technologies.

*Remark* 1. [5] The updated distribution of $\theta_1^i$ after learning $x_1^i$ is:

$$\left(\theta_1^i \left| x_1^i \right.\right) \sim N\left(\frac{\mu + \sigma^2 x_1^i}{1 + \sigma^2}, \frac{\sigma^2}{1 + \sigma^2}\right).$$

If the agents switch to a new technology , it has again the original distribution $N\left(\mu, \sigma^2\right)$. The probability of winning is:

$$\Pr\left(x^i > x^j \left| x_1^i, x_1^j \right.\right) = \Pr\left(\Delta\theta_2 < \Delta x_1 \left| \Delta x_1 \right.\right) = F_{\Delta\theta_2}\left[\Delta x_1\right],$$

where $\Delta\theta_2 = \theta_2^j - \theta_2^i$[6]. The agents choose S over K if this gives a higher probability of winning:

$$F_{\left(\theta_2^j - \theta_S^i\right)}\left(\Delta x_1\right) \geq F_{\left(\theta_2^j - \theta_K^i\right)}\left(\Delta x_1\right)$$

It will be shown, that optimal decision depends on the action of the competitor, agent $j$.

Given agent $j$ chooses S. In the following we write $\theta_S^j - \theta_K^i = \Delta\theta_{KS}$ and $F_{\left(\theta_S^j - \theta_K^i\right)} = F_{KS}$, where agent $i$ chooses S and agent $j$ chooses K and analogously for other combinations. Given agent $j$ chooses S, agent $i$ prefers S as well if this gives a higher probability of winning, otherwise agent $i$ chooses K:

[5] Derivation in Appendix A.1

[6] Note the opposing order of $i$ and $j$ in $\Delta\theta_2 = \theta_2^j - \theta_2^i$ and $\Delta x_1 = x_1^i - x_1^j$.

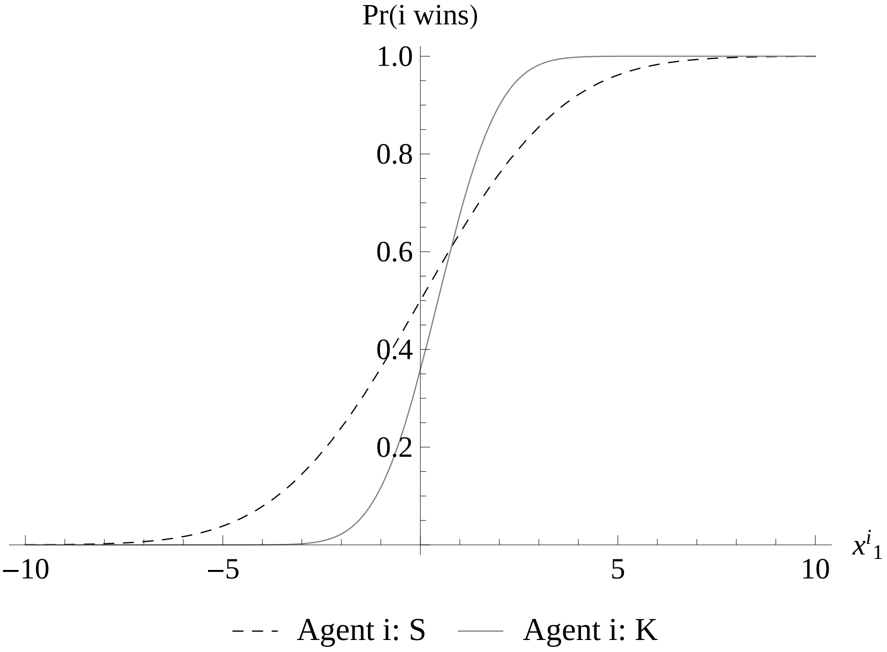

FIGURE 1. Probability of winning for agent $i$, given agent $j$ switches. (For $\mu = 1$, $\sigma = 2$, $x_1^j = 0$.)

$$\begin{aligned} F_{SS}\left[\Delta x_1\right] &\geq F_{KS}\left[\Delta x_1\right] \\ \Rightarrow \Phi\left(\frac{\Delta x_1}{\sqrt{2}\sigma}\right) &\geq \Phi\left(\frac{\Delta x_1 - \frac{\sigma^2\left(\mu - x_1^i\right)}{1+\sigma^2}}{\sigma\sqrt{1+\frac{1}{1+\sigma^2}}}\right) \end{aligned}$$

Since $\Delta\theta_{SS} \sim N\left(0, 2\sigma^2\right)$ and $\Delta\theta_{KS} \sim N\left(\frac{\sigma^2\left(\mu - x_1^i\right)}{1+\sigma^2}, \sigma^2\left(1+\frac{1}{1+\sigma^2}\right)\right)$. Otherwise agent $i$ prefers to keep the tried technology. The respective probabilities of winning can be seen in figure 1.

Since $\Phi$, the cumulative distribution function (CDF) of the standard normal distribution is monotonically increasing, this is equivalent to:

$$\begin{aligned} \Rightarrow \frac{\Delta x_1}{\sqrt{2}\sigma} &\geq \frac{\Delta x_1 - \frac{\sigma^2\left(\mu - x_1^i\right)}{1+\sigma^2}}{\sigma\sqrt{1+\frac{1}{1+\sigma^2}}} \\ \Rightarrow x_1^i &\leq \frac{x_1^j + \mu\left(2+\sqrt{2+\frac{2}{1+\sigma^2}}\right)}{3+\sqrt{2+\frac{2}{1+\sigma^2}}} \end{aligned}$$

This gives a critical value for the first period output of agent $i$, $x_1^{iS*}\left(x_1^j\right) = \frac{x_1^j+\mu\left(2+\sqrt{2+\frac{2}{1+\sigma^2}}\right)}{3+\sqrt{2+\frac{2}{1+\sigma^2}}}$, conditional on agent $j$ switching. Below $x_1^{iS*}\left(x_1^j\right)$ agent $i$ prefers S while above they prefer K. The gradient of $x_1^{iS*}\left(x_1^j\right)$ is $\frac{1}{5}$ for $\sigma = 0$ and the limit is $\frac{1}{3+\sqrt{2}}$ for $\sigma \to \infty$.

Given agent $j$ chooses K. Given agent $j$ chooses K, agent $i$ prefers S if:

$$\begin{aligned} F_{SK}\left[\Delta x_1\right] &\geq F_{KK}\left[\Delta x_1\right] \\ \Rightarrow \Phi\left(\frac{\Delta x_1 - \frac{\sigma^2\left(x_1^j-\mu\right)}{1+\sigma^2}}{\sigma\sqrt{1+\frac{1}{1+\sigma^2}}}\right) &\geq \Phi\left(\frac{\Delta x_1 - \frac{\sigma^2\left(x_1^j-x_1^i\right)}{1+\sigma^2}}{\sqrt{2\frac{\sigma^2}{1+\sigma^2}}}\right) \\ \Rightarrow x_1^i &\leq \frac{x_1^j\left(1+2\sigma^2\right)+\mu\left[2+\sqrt{2\left(2+\sigma^2\right)}\right]}{3+2\sigma^2+\sqrt{2\left(2+\sigma^2\right)}} \end{aligned}$$

Since $\Delta\theta_{SK} \sim N\left(\frac{\sigma^2\left(x_1^j-\mu\right)}{1+\sigma^2}, \sigma^2\left(1+\frac{1}{1+\sigma^2}\right)\right)$ and $\Delta\theta_{KK} \sim N\left(\mu - \frac{\mu+\sigma^2 x_1^i}{1+\sigma^2}, \sigma^2 + \frac{\sigma^2}{1+\sigma^2}\right)$. Otherwise agent $i$ prefers to keep the tried technology. This gives a critical value for the first period output of agent $i$, $x_1^{iK*}\left(x_1^j\right) = \frac{x_1^j\left(1+2\sigma^2\right)+\mu\left[2+\sqrt{2\left(2+\sigma^2\right)}\right]}{3+2\sigma^2+\sqrt{2\left(2+\sigma^2\right)}}$, conditional on agent $j$ keeping. Below $x_1^{iK*}\left(x_1^j\right)$ agent $i$ prefers S while above they prefer K.

The gradient of $x_1^{iK*}\left(x_1^j\right)$ is $\frac{1}{5}$ for $\sigma = 0$ and the limit is 1 for $\sigma \to \infty$. At the point $x_1^i = x_1^j = \mu$ we have $x_1^{iS*} = x_1^{iK*}$, so both agents are indifferent between S and K at this point.

Equilibrium. All four critical value functions can be seen in figure 2.

In equilibrium, agent $i$ prefers S if :

$$x_1^i \leq \begin{cases} x_1^{iS*}\left(x_1^j\right) & \text{and } x_1^j \leq x_1^{jS*}\left(x_1^j\right) \text{ (agent j switches and expects i to switch)} \\ x_1^{iK*}\left(x_1^j\right) & \text{and } x_1^j > x_1^{jS*}\left(x_1^j\right) \text{ (agent j keeps and expects i to switch).} \end{cases}$$

Agent $i$ prefers K if:

$$x_1^i > \begin{cases} x_1^{iS*}\left(x_1^j\right) & \text{and } x_1^j \leq x_1^{jK*}\left(x_1^j\right) \text{ (agent j switches and expects i to keep)} \\ x_1^{iK*}\left(x_1^j\right) & \text{and } x_1^j > x_1^{jK*}\left(x_1^j\right) \text{ (agent j keeps and expects i to keep).} \end{cases}$$

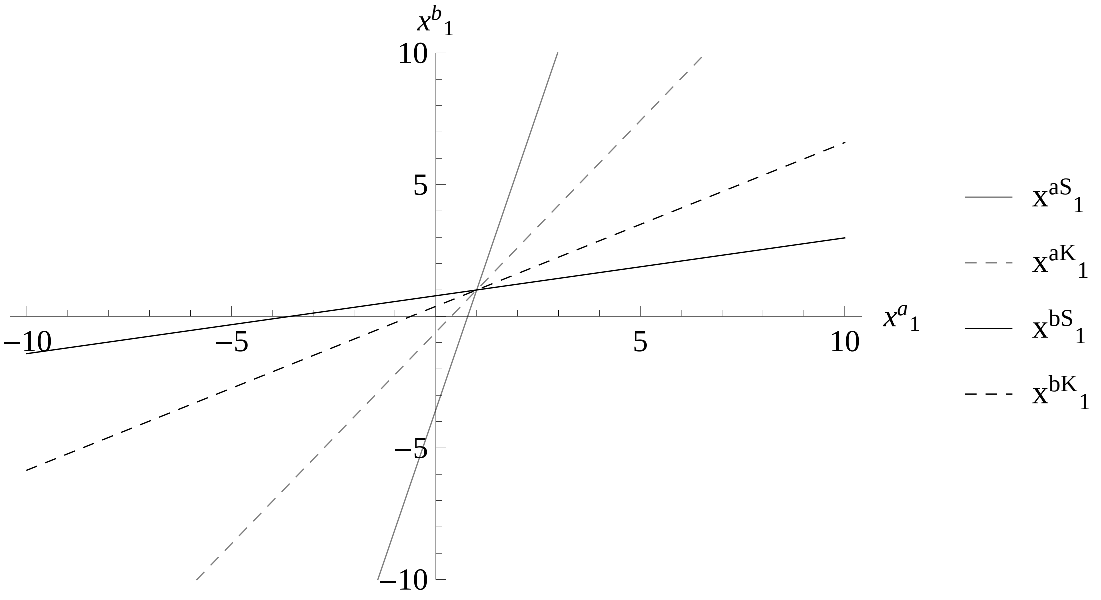

Figure 2. All critical values. (For $\mu = 1$, $\sigma = 2$.)

This equilibrium is unique, since gradients of $x_1^{iS*}\left(x_1^j\right), x_1^{iK*}\left(x_1^j\right)$ are less or equal to 1 for $\sigma\epsilon\left(0,\infty\right)$, such that the order never changes.[7]

In equilibrium, the critical values can be simplified to:

$$x_1^{i*}\left(x_1^j\right) = \begin{cases} x_1^{iS*}\left(x_1^j\right) & \text{if } x_1^j \leq \mu \\ x_1^{iK*}\left(x_1^j\right) & \text{if } x_1^j \leq \mu. \end{cases}$$

Thus, agent $i$ prefers S if $x_1^i \leq x_1^{i*}$ and K otherwise. The resulting relevant critical value functions for both agents can be seen in figure 3.

In order to maximise output, agent $i$ should switch when output is below the expectation ($x_1^i \leq \mu$) and keep the tested technology when it's above, independently from the output of agent $j$. Hence, the behaviour of the agents is not optimal from the perspective of the principal. The divergence of the blue line from a horizontal line and of the purple line of a vertical line through $\mu$ in figure 3 represents this inefficiency.

Figure 4 shows the regions of inefficient switching for agent $i$. In region A agent $i$ inefficiently chooses K, even though output is below the expectation $\mu$. This is

[7] Specifically, where $x_1^i\epsilon\left(x_1^{iS*}\left(x_1^j\right), x_1^{iK*}\left(x_1^j\right)\right)$, i.e. the decision between S and K depends on the decision of the competitor, we always have either that the competitor would choose S $\left[x_1^j < x_1^{jS*}\left(x_1^j\right) \text{ and } x_1^j < x_1^{jK*}\left(x_1^j\right)\right]$ or respectively choose K $\left[x_1^j > x_1^{jS*}\left(x_1^j\right) \text{ and } x_1^j > x_1^{jK*}\left(x_1^j\right)\right]$ independently of the decision of agent $i$. Hence, the equilibrium is unique.

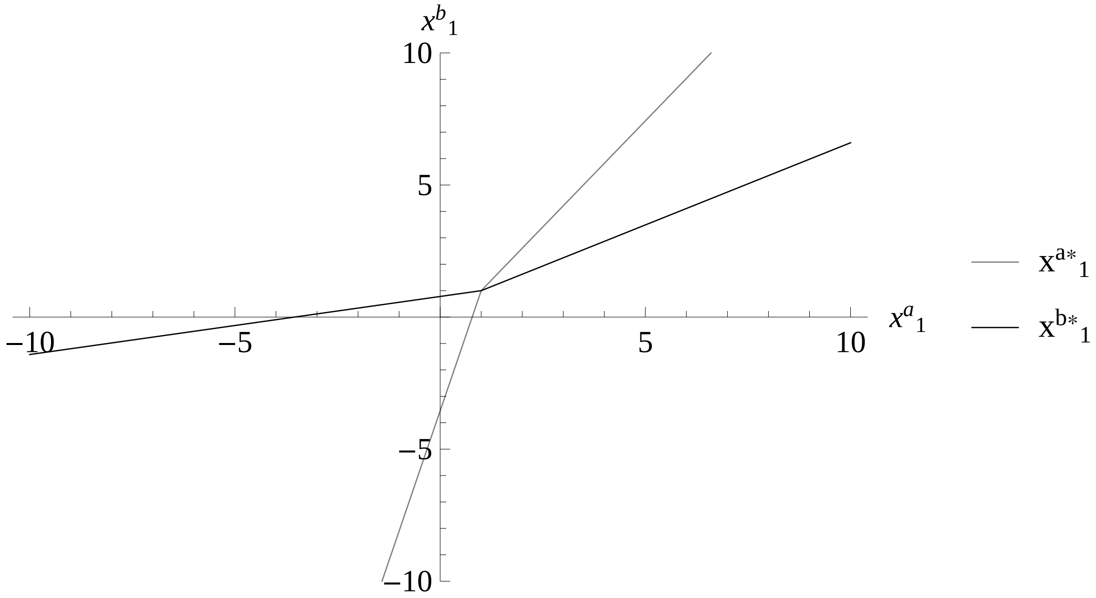

FIGURE 3. Relevant critical value functions. (For $\mu = 1$, $\sigma = 2$.)

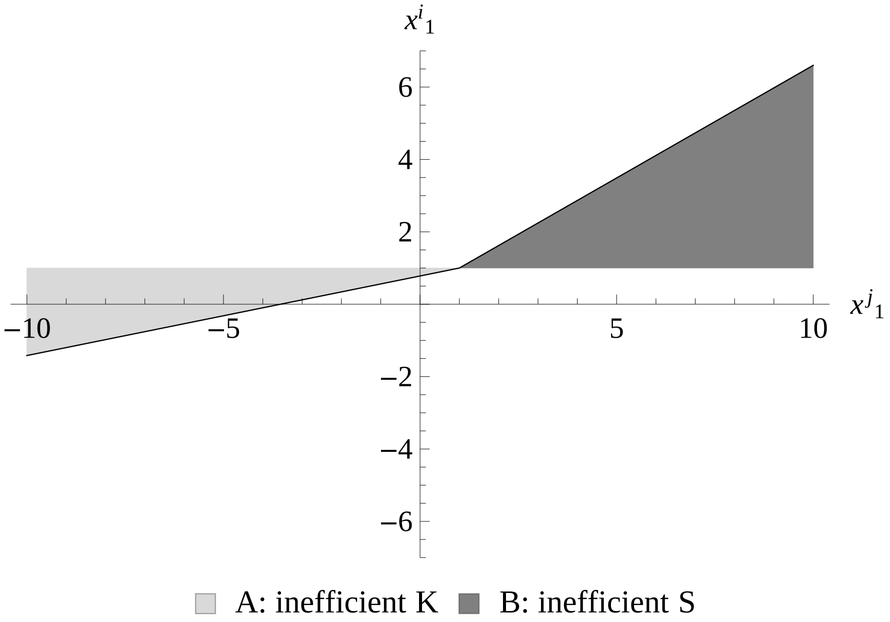

FIGURE 4. Inefficient switching decisions of agent $i$. (For $\mu = 1$, $\sigma = 2$.)

because in this region agent $j$'s output is even worse (since $\frac{\partial x_1^{iS*}(x_1^j)}{\partial x_1^j} \leq 1$). Agent $i$ prefers to keep the old technology which has a lower mean, but also a smaller variance after updating than a new technology would have. Since agent $i$ has already an output advantage over $j$, the probability of winning with a technology with a below-average but safer output is higher than with a new, more variable technology. In region B, on the other hand, agent $i$ inefficiently chooses S, even though output from the period 1 technology is above the expectation $\mu$. This is because in this region agent $j$'s output is even better (since $\frac{\partial x_1^{iK*}(x_1^j)}{\partial x_1^j} \leq 1$). Agent $i$ prefers to try a new technology which has a lower mean, but also a higher variance. Since agent $i$ is already lagging behind $j$ in terms of output, the probability of winning is higher if agent $i$ takes a chance on a new technology which might be more productive. Even though the agents are risk-neutral, the structure of the tournament lets agent a act similar to a risk-averse agent in region A and to a risk-loving agent in region B. However, even though the agents' switching decisions are not optimal from the perspective of the principal, the outcome is superior to the case with no feedback where the agents always choose S. This would be suboptimal in the whole region where $x_1^j > \mu$

The next section will show that the principal can reduce this inefficiency by assigning different weights on period 1 and 2.

## 3. Optimal Allocation

In this section the assumption that the prize goes to the agent with the higher aggregate output is relaxed. Instead, the principal decides how to weigh the two periods when allocating the prize. In this setting the prize goes to the agent with the higher weighted output: $x^i = r\, x_1^i + (1-r)\; x_2^i$ and the principal optimally sets the weight $r$, with $r \epsilon\, [0,1)$. Again, the agents decide between S and K in order to maximize the probability of winning which is $\Pr\left(r\, x_1^i + (1-r)\; \theta_2^i > r\, x_1^j + (1-r)\; \theta_2^j\right)$ in this case. If the agents do not get any feedback the situation is as in section 2 and we can assume them to always choose S. If the principal gives feedback and publicly reveals $x_1^a, x_1^b$ the agents choose between S and K to maximise the probability of winning:

$$\begin{aligned} &\Pr\left(r\, x_1^i + (1-r)\; \theta_2^i > r\, x_1^j + (1-r)\; \theta_2^j \,\middle|\, x_1^a, x_1^b\right) \\ =&F_{\Delta\theta_2}\left[\frac{r}{1-r}\,\Delta x_1\right] \\ =&\Phi\left[\frac{\frac{r}{1-r}\Delta x_1 - \mu_{\Delta\theta_2}}{\sigma_{\Delta\theta_2}}\right], \end{aligned}$$

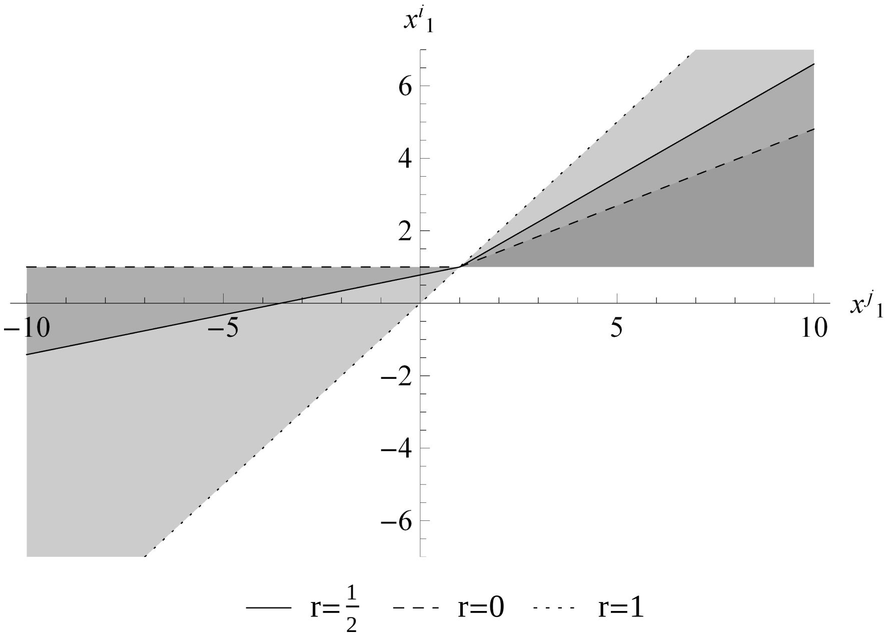

Figure 5. Inefficient switching decisions of agent $i$ for different weights. (For $\mu = 1$, $\sigma = 2$.)

where $\Delta\theta_2 \sim N\left(\mu_{\Delta\theta_2}, \sigma_{\Delta\theta_2}\right)$.

Analogously to section 2 there is a unique equilibrium where agent $i$ prefers S if $x^i_1 \leq \hat{x}^i_1\left(x^j_1, r\right)$ and K otherwise, where the critical value $\hat{x}^i_1\left(x^j_1, r\right)$ is defined as:

$$\hat{x}^i_1\left(x^j_1, r\right) = \begin{cases} \hat{x}^{iS}_1\left(x^j_1\right) & \text{if } x^j_1 \leq \mu \\ \hat{x}^{iK}_1\left(x^j_1\right) & \text{if } x^j_1 \leq \mu \end{cases}$$

and where:

$$\begin{aligned} \hat{x}^{iS}_1\left(x^j_1\right) &= \frac{\frac{r}{1-r}\, x^j_1 + \mu\left(2 + \sqrt{2 + \frac{2}{1+\sigma^2}}\right)}{\frac{r}{1-r} + 2 + \sqrt{2 + \frac{2}{1+\sigma^2}}} \\ \hat{x}^{iK}_1\left(x^j_1\right) &= \frac{x^j_1\left[\frac{r+\sigma^2}{1-r}\right] + \mu\left(2 + \sqrt{2\left(2+\sigma^2\right)}\right)}{\frac{r+\sigma^2}{1-r} + 2 + \sqrt{2\left(2+\sigma^2\right)}}. \end{aligned}$$

The critical value $\hat{x}^i_1\left(x^j_1, r\right)$can be seen in figure 5 for different weights $r$.

**For** $r = \frac{1}{2}$

where *both periods carry equal weight*, we naturally have $\hat{x}^i_1 = x^{i*}_1$, i.e. the critical value is equal to the one derived in section 2.

**For** $r = 0$

where *only period 2 counts*, we have $\hat{x}_1^{iS} = \mu$. Thus, agents make the efficient switching decision in the area where the competitor chooses S. If period 1 does not affect the allocation there is no inefficient keeping, because the agents cannot gain a head start. However, agents do not make the efficient switching decision in the area where the competitor chooses K. Here the critical value becomes $\hat{x}_1^{iK}\left(x_1^j\right) = \frac{\sigma^2 x_1^j + \mu\left(2+\sqrt{2(2+\sigma^2)}\right)}{\sigma^2+2+\sqrt{2(2+\sigma^2)}}$. The smaller the variance of the technology $\sigma^2$, the closer to efficiency the switching decision becomes, with $\lim_{\sigma \to 0} \hat{x}_1^{iK} = \mu$. [8] On the other hand inefficiency increases as $\sigma$ grows, with $\lim_{\sigma \to \infty} \hat{x}_1^{iK} = x_1^j$. This is the most inefficient behaviour, since the switching decision of agent $i$ does not depend on their own output $x_1^j$ at all any more, but only on the competitor's output $x_1^j$. Thus, also when all weight is on period 2, there is inefficient switching, though a little less than in the case with equal weights. This is because, in the concerned area agent $i$'s first period output $x_1^j$ is above expectation, but below the competitor's output $x_1^j$. Thus the updated expectation for second period output from the tested technology for agent $i$ will be lower. Thus, even though the competitor cannot gain a head start in period 1, agent $i$ prefers to take a chance on a new technology with higher variance in order to have a higher chance to outdo the competitor's technology.

**For** $r \to 1$

where *period 2 counts very little* compared to period 1, we have $\lim_{r \to 1} \hat{x}_1^i = x_1^j$. The agents follow the most inefficient rule, independently from the switching decision of the competitor. Agent $i$ always chooses S if her output is below the competitor's, even if it is above the expectation, because they can only win if they have a huge 2nd period output. On the other hand, agent $i$ always keeps if her output is above the competitor's, even if it is below the expectation, because it is very unlikely that the competitor will catch up with them. However, if $r = 1$, i.e. when the second period output is completely irrelevant for prize allocation, agents are indifferent about switching, since the tournament is already decided after period

[8] However, for the extreme case of $\sigma = 0$ the problem is meaningless, since there is no uncertainty about the technology, which has a fixed value of $\mu$.

1. Hence, they can be assumed to follow the principal's wishes and switch when $x_1^j \leq \mu$. Therefore, the principal can achieve the efficient switching decision only by making it irrelevant for the outcome of the tournament. However, this will normally not be optimal.

In summary, the principal can indeed improve the switching behaviour of the agents by increasing the weight on period 2. Alternatively, the optimal switching behaviour can be achieved by putting all weight on period 1, such that the allocation of the prize is completely independent from the behaviour of the agents. Both alternatives might yet not be optimal in a more realistic setting where agents exert effort, since the agents would not exert any more effort in period 2. This will be explored in section 4.

## 4. Effort

This section considers the case where effort $(e_2^i \geq 0)$ enters additively in the output.[9] First period output of agent $i$ becomes $x_1^i = e_1^i + \theta_1^i + \varepsilon_1^i$ while second period output is $x_2^i = e_2^i + \theta_2^i$. The cost of effort is fixed to $c(e) = \frac{e^2}{2}$. For quadratic effort Ederer (2010) shows that the feedback policy does not affect the output level in a setting without learning about technologies. However, it will be shown that this is not the case when agents can switch technologies. As in section 3, the prize goes to the agent with the higher weighted output: $x^i = r\, x_1^i + (1-r)\, x_2^i$, with $r \epsilon [0,1)$. The optimal effort and switching decision can be derived using backwards induction. Agents exert higher effort after receiving feedback and - predictably - they exert less effort in a period that has less weight for the allocation of the prize. First, the case where the agents receive full feedback is considered.

### 4.1. **Full Feedback.**

4.1.1. *Second Period Effort.* After learning $x_1^i, x_1^j$ and choosing $a^i \epsilon \{S, K\}$, agent $i$ chooses effort $e_2^i$ to maximise expected utility $EU_2^i$, given by the probability of winning minus the cost of effort.

[9] It would also be very interesting and perhaps more realistic to consider the case where effort enters as a multiplier of productivity. However the problem becomes algebraically very complex and it is not possible to get a general analytic solution.

$$\begin{aligned} EU_2^i\left(e_2^i\right) =& \Pr\left[r\left(e_1^i+\theta_1^i+\varepsilon_1^i\right)+(1-r)\left(e_2^i+\theta_2^i\right)\right.\\ &\left.> r\left(e_1^j+\theta_1^j+\varepsilon_1^j\right)+(1-r)\left(e_2^j+\theta_2^j\right)\left|x_1^a,x_1^b,a^i\right.\right]-c\left(e\right)\\ =&F_{\Delta\theta_2}\left[\frac{r}{1-r}\Delta x_1+e_2^i-e_2^j\left|x_1^i,x_1^j,a^i\right.\right]-\frac{\left(e_2^i\right)^2}{2} \end{aligned}$$

The first order condition (FOC) is:

$$e_2^{iF}=f_{\Delta\theta_2}\left(\frac{r}{1-r}\Delta x_1+e_2^i-e_2^j\left|x_1^a,x_1^b,a^i\right.\right)$$

Here, and in the remainder of the section, the analysis is restricted to cases where the FOC is sufficient to determine the optimal effort.[10] Then there is a unique symmetric equilibrium where both agents exert the same effort $\left(e_2^i=e_2^j\right)$, since the FOC have the same value for both agents also when one is leading in terms of first period output:

$$f_{\Delta\theta_2}\left[\frac{r}{1-r}\Delta x_1+e_2^i-e_2^j\left|x_1^a,x_1^b,a^i\right.\right]=f_{(-\Delta\theta_2)}\left[\frac{r}{1-r}\Delta x_1+e_2^j-e_2^i\left|x_1^a,x_1^b,a^i\right.\right].$$

Because the distribution is symmetric, we have $f_{\Delta\theta_2}\left(z\left|x_1^i,x_1^j,a^i\right.\right)=f_{(-\Delta\theta_2)}\left(-z\left|x_1^i,x_1^j,a^i\right.\right)$ for all first period outputs and switching decisions. Hence, in equilibrium we have:

$$e_2^{iF}=e_2^{jF}=e_2^F=f_{\Delta\theta_2}\left[\frac{r}{1-r}\left(x_1^i-x_1^j\right)\left|x_1^a,x_1^b,a^i\right.\right] \qquad (1)$$

Since $\lim\limits_{r\to1}\frac{r}{1-r}\left(x_1^i-x_1^j\right)=\infty$, we also have $\lim\limits_{r\to1}e_2^F=0$ for all switching decisions. Second period output $e_2^F$ goes to zero as $r$ approaches 1. Clearly, the agents will not exert any effort in period 2 if second period output does not count for the prize.

4.1.2. *Switching decisions.* Before choosing second period effort $e_2^i$ and after learning the first period output $x_1^a,x_1^b$, agent $i$ decides between S and K. The expected payoff is:

[10] $EU_2^i\left(e_2^i\right)$ is concave, however for small values of $\sigma$ the effort level $e_2^{i*}$ given by the FOC can become larger than the resulting probability of winning. In that case, there would be no pure strategy equilibrium. These cases are excluded from consideration here.

$$\begin{aligned}
&F_{\Delta\theta_2}\left[\frac{r}{1-r}\,\Delta x_1\left|x_1^a, x_1^b\right.\right]-c\left(e_2^{iF}\right)\\
=&F_{\Delta\theta_2}\left[\frac{r}{1-r}\,\Delta x_1\left|x_1^a, x_1^b\right.\right]-\frac{1}{2}\left\{f_{\Delta\theta_2}\left[\frac{r}{1-r}\,\Delta x_1\left|x_1^a, x_1^b\right.\right]\right\}^2\\
=&\Phi\left[\frac{\frac{r}{1-r}\,\Delta x_1-\mu_{\Delta\theta_2}}{\sigma_{\Delta\theta_2}}\right]-\frac{1}{2}\,\phi\left[\frac{\frac{r}{1-r}\,\Delta x_1-\mu_{\Delta\theta_2}}{\sigma_{\Delta\theta_2}}\right]^2
\end{aligned}$$

where $\Delta\theta_2 \sim N\left(\mu_{\Delta\theta_2}, \sigma_{\Delta\theta_2}\right)$. Setting $g\left(z\right)=\Phi\left(z\right)-\frac{1}{2}\,\phi\left(z\right)^2$, we find that $g\left(z\right)$ is still monotonically increasing (as is $\Phi\left(z\right)$), so the agent will make the choice that gives a higher value of $z$, just as in the case without effort. Therefore the agents have exactly same critical values for switching $\hat{x}_1^i\left(x_1^j, r\right)$ as in section 3.

4.1.3. *First Period Effort.* At the beginning of the tournament, agent $i$ chooses first period effort $e_1^i$ to maximise expected utility which takes the form:

$$\begin{aligned}
EU_2^i\left(e_2^i\right)=&\Pr\left[r\left(e_1^i+\theta_1^i+\varepsilon_1^i\right)+(1-r)\left(e_2^i+\theta_2^i\right)\right.\\
&\left.>r\left(e_1^j+\theta_1^j+\varepsilon_1^j\right)+(1-r)\left(e_2^j+\theta_2^j\right)\right]-\frac{\left(e_1^i\right)^2}{2}-E\left[\frac{\left(e_2^i\right)^2}{2}\right]\\
=&E_{\theta_2^i,\theta_2^j}F_{\left(\theta_1^j+\varepsilon_1^j-\theta_1^i-\varepsilon_1^i\right)}\left[e_1^i-e_1^j+\frac{1-r}{r}\left(\theta_2^i-\theta_2^j+e_2^i-e_2^j\right)\right]\\
&-\frac{\left(e_1^i\right)^2}{2}-E\left[\frac{\left(e_2^i\right)^2}{2}\right]
\end{aligned}$$

The FOC is:

$$\begin{aligned}
e_1^i=\quad&E_{\theta_2^i,\theta_2^j}\,f_{\left(\theta_1^j+\varepsilon_1^j-\theta_1^i-\varepsilon_1^i\right)}\left[e_1^i-e_1^j+\tfrac{1-r}{r}\left(\theta_2^i-\theta_2^j+e_2^i-e_2^j\right)\right]\\
&+E_{\theta_2^i,\theta_2^j}\left(\left\{f_{\left(\theta_1^j+\varepsilon_1^j-\theta_1^i-\varepsilon_1^i\right)}\left[e_1^i-e_1^j+\tfrac{1-r}{r}\left(\theta_2^i-\theta_2^j+e_2^i-e_2^j\right)\right]-e_2^i\right\}\tfrac{\partial e_2^i}{e_1^i}\right)\\
&-E_{\theta_2^i,\theta_2^j}\left(\left\{f_{\left(\theta_1^j+\varepsilon_1^j-\theta_1^i-\varepsilon_1^i\right)}\left[e_1^i-e_1^j+\tfrac{1-r}{r}\left(\theta_2^i-\theta_2^j+e_2^i-e_2^j\right)\right]\right\}\tfrac{\partial e_2^j(\Delta x_1)}{e_1^i}\right)
\end{aligned}$$

As above, we restrict the analysis to cases where the FOC is sufficient to determine the optimal effort. Then the equilibrium effort simplifies to:

$$e_1^F = E_{\theta_2^i,\theta_2^j} f_{\left(\theta_1^j+\varepsilon_1^j-\theta_1^i-\varepsilon_1^i\right)} \left[\frac{1-r}{r}\left(\theta_2^i - \theta_2^j\right)\right], \tag{2}$$

Also first period effort is symmetric, due to the symmetry of $f_{\left(\theta_1^j+\varepsilon_1^j-\theta_1^i-\varepsilon_1^i\right)}$. The second and third term can be shown to equal zero.[11] Thus in equilibrium, first period effort does not have a strategic component affecting second period output of either agent. We have$\left(\theta_1^j + \varepsilon_1^j - \theta_1^i - \varepsilon_1^i\right) \sim N\left(0, 2+2\sigma^2\right)$. Since $\lim\limits_{r\to 0}\frac{1-r}{r}\left(\theta_2^i - \theta_2^j\right) = \infty$, we also have $\lim\limits_{r\to 0} e_1^{iF} = 0$, first period output $e_1^{iF}$ goes to zero as $r$ approaches zero. Clearly, the agents will not exert any effort in period 1 if first period output does not count for the prize.

4.2. **No Feedback.** Since the principal does not give feedback, the agents cannot condition their switching decisions on first period output. As stated in the basic model, both switching and keeping would give agents an equal probability of winning. Thus, in the model without effort they are indifferent. The switching decision will, however, affect equilibrium effort. Specifically, if agent $i$ keeps the technology, aggregate productivity has a higher variance, independent of the switching decision of the competitor. The value of $f_{\left(\theta_1^j+\theta_2^j-\theta_1^i-\theta_2^i\right)}(0)$, and therefore the equilibrium level of effort, will be lower. This is because for a higher variance of the aggregate output, the effort is expected to make less of a difference in the probability of winning. Consequently, the agents prefer to keep the technology since this will lead to lower effort and lower cost for her. So without feedback agents keep their technology.

4.2.1. *Second Period Effort.* The symmetric equilibrium effort in the second period will be:

[11] Proof analogous to the proof in appendix A.1 in Ederer (2010).

$$
\begin{aligned}
e_2^N =& E_{\Delta x_1}\left\{f_{\left(\theta_2^j-\theta_2^i\right)}\left[\frac{r}{1-r}\left(\Delta x_1\right)\right]\right\}\\
=& \int\limits_{-\infty}^{\infty} f_{\Delta x_1}\left(\Delta x_1\right)\, f_{KK}\left[\frac{r}{1-r}\left(\Delta x_1\right)\right]\, d\Delta x_1, \qquad (3)
\end{aligned}
$$

where $\Delta x_1 \sim N\left(0, 2\sigma^2+2\right)$, since first period effort will turn out to be symmetric.

4.2.2. *First Period Effort.* The symmetric equilibrium effort in the first period has the same form as in the case with full feedback:

$$
\begin{aligned}
e_1^N =& E_{\theta_2^i,\theta_2^j}\, f_{\left(\theta_1^j+\varepsilon_1^j-\theta_1^i-\varepsilon_1^i\right)}\left[\frac{1-r}{r}\left(\theta_2^i-\theta_2^j\right)\right]\\
=& E_{\Delta\theta_2}\, f_{\left(\theta_1^j+\varepsilon_1^j-\theta_1^i-\varepsilon_1^i\right)}\left[\frac{1-r}{r}\left(-\Delta\theta_2\right)\right]\\
=& \int\limits_{-\infty}^{\infty} f_{KK}\left(\Delta\theta_2\right)\, f_{\Delta x_1}\left[\frac{1-r}{r}\left(\Delta\theta_2\right)\right]\, d\Delta\theta_2, \qquad (4)
\end{aligned}
$$

since $\left(\theta_1^j+\varepsilon_1^j-\theta_1^i-\varepsilon_1^i\right) \sim N\left(0, 2\sigma^2+2\right)$. Thus, for $r=\frac{1}{2}$, when period 1 and 2 carry the same weight, we have that effort in both periods is equal: $e_1^N = e_2^N$. This is intuitive, given increasing marginal cost of effort and the absence of additional information between the periods. This is not the case when different weights are assigned to period 1 and 2. Clearly, for $r > \frac{1}{2}$, first period output will be higher and for $r < \frac{1}{2}$ second period output will be higher.

4.3. **Comparison**[12]**.** Second period effort with feedback $e_2^F$ as given by 1 will vary depending on the values of $x_1^a, x_1^b$. The expectation of $e_2^F$ before the first period is $E_{\Delta x_1}\left(e_2^F\right) = \int_{-\infty}^{\infty} f_{\Delta x_1}\left(\Delta x_1\right)\, f_{\theta_2}\left[\frac{r}{1-r}\left(\Delta x_1\right)\right]\, d\Delta x_1$. The only difference to the second period effort without feedback $e_2^N$, as given by 3, is the distribution of second period technologies $f_{\theta_2}$. This will be different from $f_{KK}$ since the agents will not always switch, but base their switching decision on the feedback they receive. The resulting distribution $f_{\theta_2}$ will have a mean of zero, due to the symmetry of

[12] This part still needs to be revised. I'm working on a proof.

the distributions, but it has a lower variance. Thus, the expected effort level with feedback is higher than without in both periods.

## 5. Partial Feedback: Recommendations

While in the previous sections the principal was restricted to either giving full public feedback or no feedback at all, this assumption is now relaxed such that the principal can also give partial and private feedback. Specifically, the case is considered where the principal *privately* sends a message $m \epsilon (S, K)$ to each agent, recommending them S if $x_1^i \leq \mu$ and K if $x_1^i > \mu$. It will be shown, that this policy is indeed incentive compatible. The agents prefer to follow the recommendation, given the information conveyed and the efficient switching behaviour is implemented.

### 5.1. Principal recommends S.

If the principal privately recommends S to agent $i$ when $x_1^i \leq \mu$, the ex post probability of winning is:

$$\begin{aligned} & \Pr\left(x_1^i + \theta_2^i > x_1^j + \theta_2^j \,|x_1^i \leq \mu\right) \\ = \; & 1 - E_{x_1^j, \theta_2^j, x_1^i}\left[F_{\theta_2^i}\left(x_1^j + \theta_2^j - x_1^i \,|x_1^i \leq \mu\right) |x_1^i \leq \mu\right] \end{aligned}$$

**Proposition 1.** *If the principal privately recommends S to agent i if $x_1^i \leq \mu$, agent i prefers to follow the recommendation.*

Agent $i$ will follow recommendation of S if this gives a higher probability of winning than K, given that $x_1^i \leq \mu$. For this to be the case we need:

$$\begin{aligned} & 1 - E_{x_1^j, \theta_2^j, x_1^i}\left[F_{\theta_S^i}\left(x_1^j + \theta_2^j - x_1^i \left|x_1^i \leq \mu\right.\right) \left|x_1^i \leq \mu\right.\right] \\ \geq & 1 - E_{x_1^j, \theta_2^j, x_1^i}\left[F_{\theta_K^i}\left(x_1^j + \theta_2^j - x_1^i \left|x_1^i \leq \mu\right.\right) \left|x_1^i \leq \mu\right.\right] \end{aligned}$$

$$\begin{aligned}\Rightarrow & E_{x_1^j,\theta_2^j,x_1^i}\left[F_{\theta_K^i}\left(x_1^j+\theta_2^j-x_1^i\left|x_1^i\leq\mu\right.\right)\right.\\ & \left.-F_{\theta_S^i}\left(x_1^j+\theta_2^j-x_1^i\left|x_1^i\leq\mu\right.\right)\left|x_1^i\leq\mu\right.\right]\geq 0\end{aligned}$$

A sufficient (but not necessary) condition for this to hold is:

$$F_{\theta_K^i}\left(z\left|x_1^i\leq\mu\right.\right)-F_{\theta_S^i}\left(z\left|x_1^i\leq\mu\right.\right)\geq 0\quad\forall z, \tag{5}$$

i.e. $F_{\theta_K^i}\left(z\left|x_1^i\leq\mu\right.\right)$ first order stochastically dominates $F_{\theta_S^i}\left(z\left|x_1^i\leq\mu\right.\right)$. The proof that 5 holds is in appendix A.2.

### 5.2. **Principal recommends K.**

**Proposition 2.** *If the principal privately recommends K to agent $i$ if $x_1^i>\mu$, agent $i$ prefers to follow the recommendation.*

If the principal privately recommends K to agent $i$ when $x_1^i>\mu$, the agent will follow if K gives a higher probability of winning than S:

$$\begin{aligned}\Rightarrow & E_{x_1^j,\theta_2^j,x_1^i}\left[F_{\theta_S^i}\left(x_1^j+\theta_2^j-x_1^i\left|x_1^i\leq\mu\right.\right)\right.\\ & \left.-F_{\theta_K^i}\left(x_1^j+\theta_2^j-x_1^i\left|x_1^i\leq\mu\right.\right)\left|x_1^i\leq\mu\right.\right]\geq 0\end{aligned}$$

A sufficient (but not necessary) condition for this to hold is:

$$F_{\theta_S^i}\left(z\left|x_1^i>\mu\right.\right)-F_{\theta_K^i}\left(z\left|x_1^i>\mu\right.\right)\geq 0\,\forall z \tag{6}$$

i.e. $F_{\theta_S^i}\left(z\left|x_1^i>\mu\right.\right)$ first order stochastically dominates $F_{\theta_K^i}\left(z\left|x_1^i>\mu\right.\right)$. The proof that 6 holds is in appendix A.3.

Since both propositions hold, the efficient switching decision can be implemented with a private partial feedback scheme. If the principal privately recommends S to agent $i$ if $x_1^i\leq\mu$ and K if $x_1^i>\mu$, it is in the agents' best interest to follow the recommendation.

5.3. **The Case of n Agents with Partial Feedback.** One might think that recommendations as characterized above will not be able to implement the efficient switching behaviour when the number of agents is increased. If the agents have to face more and more competitors, it can be expected that whoever is the leader after period 1 has an increasingly high output. So it might seem that always choosing S in order to have a chance at catching the leader becomes more attractive - as in the case inefficient switching in the setting with full feedback in order to catch a far-ahead competitor. However it can be shown that this is actually not the case. In a setting with an arbitrary number of competitors, each agent still wants to follow the recommendation of the principal if the principal follows the same policy as above, recommending S for $x_1^i \leq \mu$ and K for $x_1^i > \mu$.

In a setting with $n$ agents, the probability of winning for agent $i$ given message $m$ is:

$$\begin{aligned} & \Pr\left(x_1^i + \theta_2^i > x^j \left| m\right.\right) \forall j \neq i \\ = \; & E_{x^j, x_1^i}\left\{\prod_{j \neq i}\left[1 - F_{\theta_2^i}\left(x^j - x_1^i \left| m\right.\right)\right] \left| m\right.\right\} \end{aligned}$$

If the principal privately recommends S to agent $i$ when $x_1^i \leq \mu$, the agent follows if S gives a higher probability of winning than K:

$$\begin{aligned} & E_{x^j, x_1^i}\left\{\prod_{j \neq i}\left[1 - F_{\theta_S^i}\left(x^j - x_1^i \left| m\right.\right)\right] \left| m\right.\right\} \\ \geq & E_{x^j, x_1^i}\left\{\prod_{j \neq i}\left[1 - F_{\theta_K^i}\left(x^j - x_1^i \left| m\right.\right)\right] \left| m\right.\right\} \end{aligned}$$

$$\begin{aligned} \Rightarrow & E_{x^j, x_1^i}\left\{\prod_{j \neq i}\left[1 - F_{\theta_S^i}\left(x^j - x_1^i \left| m\right.\right)\right]\right. \\ & \left. - \prod_{j \neq i}\left[1 - F_{\theta_K^i}\left(x^j - x_1^i \left| m\right.\right)\right] \left| m\right.\right\} \end{aligned} \tag{7}$$

A sufficient condition for 7 to hold is still equation 5:

$$F_{\theta_K}\left(z\left|x_1 \leq \mu\right.\right) - F_{\theta_S}\left(z\left|x_1 \leq \mu\right.\right) \geq 0 \quad \forall z.$$

The case for a recommendation of K when $x_1^i > \mu$ is analogous. Therefore, the agent wants to follow the recommendation of the principal in the case with $n$ competitors just as in the case with only 2. The difference in probability between following the recommendation and deviating becomes smaller, since the probability of winning in both cases decreases as the number of agents grows. However, the relation between the probabilities is never reversed. Hence, it can be concluded that agents follow the recommendation of the principal and the efficient switching behaviour can be achieved, independently of number of competitors.

## 6. Conclusions

This paper has studied a tournament setting where agents require feedback to learn about the productivity of their chosen technology. It has been shown that full information revelation by the principal does not lead to an optimal technology choice by the agents. Due to the tournament structure, the agents care about winning instead of maximizing output. There is inefficient switching, where an agent changes an above-average technology in order to have a better chance to outdo the competitor. On the other hand, there is inefficient keeping, where an agent holds on to a below-average technology if it is sufficient to beat the competitor. Furthermore, it is found that this inefficiency can be ameliorated when all weight is put on second period output for the allocation of the prize. However, this will likely not be optimal in a setting when agents have to exert effort. On the other hand, the efficient technology choice can be achieved if the principal is able to give partial feedback in the form of recommendations. If the principal does not reveal the output, but only recommends the agents to switch or keep technology when it is efficient to do so, it is in the best interest of the agents to follow the recommendation.

## Chapter 2. The Kindness of Strangers

### 1. Introduction

> *Last night, after donating the last of my change to Children In Need (a UK telethon appeal), I got on the train from London to Manchester. Feeling hungry, I went to the buffet car, only to find that the card machine was broken and I couldn't buy a sandwich. I turned to walk back to my seat without anything to eat, but was stopped by the man behind me who paid for the things I had tried to purchase. It was a spontaneous act of kindness from a complete stranger and left me with a great feeling. Thank you to all those people out there who try in small ways to make the world a better place!* Anonymous post, 23/6/2012.[13]

Throughout modern history, thinkers of the most diverse backgrounds considered kindness as one of the highest peaks touched by mankind. Jean-Jacques Rousseau asked, "What wisdom can you find that is greater than kindness?" The Dalai Lama said: "There is no need for temples, no need for complicated philosophies. My brain and my heart are my temples; my philosophy is kindness." William Wordsworth wrote, "The best portion of a good man's life is his little, nameless, unremembered acts of kindness and of love." Aldous Huxley maintained, "It is a bit embarrassing to have been concerned with the human problem all one's life and find at the end that one has no more to offer by way of advice than 'try to be a little kinder."'

As shown in a UK survey conducted by Griffith et al. (2011), also ordinary people care greatly about kindness: for many —as for the famous thinkers—, it is the single most important contributor to their quality of life. Moreover, the researchers document, experience of unkindness can exert an even greater influence on people's perception of social health than crime statistics.

[13] Retrieved the 30/7/2012 from http://www.helpothers.org/story.php?sid=31804. Numerous stories of kindness are uploaded anonymously on a daily basis on many websites close in spirit to helpothers.org, check for instance: randomactsofkindness.org, thekindnessofstrangers.net, and politestranger.com.

In the current paper, we propose a theoretical framework to study acts of kindness, *from* and *to* strangers, along the lines of the story reported above. In the following, we define an act of kindness as helpful behaviour towards people in need which are *not* enforceable. We say that a society of strangers has the *social norm of kindness* if its members perform, whenever in their power, acts of kindness. We stylize the relevant situation where an act of kindness can arise as a dictator game played among a pair of strangers randomly matched from some population. Since —as Bardsley (2008) claims—, everyone faces dictator games all day, every day, the stage game is repeated and at each round every person is randomly rematched in a new pair. The 'roles' of the strangers (who the 'helper' is and who the 'receiver' is) are randomly assigned at the beginning of each stage game. The possibility of *switching* role across different stage games —in the story above: of being sometimes without enough cash and sometimes with five extra pounds in the wallet during different train journeys—, captures the idea that "life is like a wheel." This layer of risk about one's future role in the dictator games to come is precisely what drives our results: entirely *selfish* people can help today some strangers in need in order to fuel the social norm of kindness, so as to increase the likelihood of receiving some help, if needed, in the future.[14] In other words, we propose to interpret 'kindness from strangers' as an *indirectly* reciprocal outcome[15] driven by a *selfish* motive. In our model, as Sophocles would say —and Adam Smith would reiterate—, "kindness is ever the begetter of kindness."

This intuitive mechanism represents an attempt of formalization of the evidences gathered by Griffith et al. (2011). The researchers document that people's understanding of kindness is in terms of how they would like to be treated: a man from Wiltshire described it as "Treating people how you would like to be treated yourself." Those interviewed tend to stress the reciprocal character of kindness:

[14] We do not study the case of non-verifiable roles. Throughout the paper, an agent's role is publicly observable.

[15] *Directly* reciprocal outcomes arise when the *same* two agents interact repeatedly: A helps B and then B helps A. *Indirectly* reciprocal outcomes arise when *different* agents interact repeatedly: A helps B, B helps C, C helps D, and so on. For a broad discussion on reciprocity see, for instance, Nowak & Sigmund (2005). For a recent survey on indirect reciprocity, see Sigmund (2012) and the references therein.

"We give out and we get back," as a taxi driver puts it. Furthermore, the researchers underline, the social norm of kindness requires to be constantly fed in order to survive: "The old people do not respect us, so we wind them up. Why not? They started it," a teenage mother complains.

> *I was in Toronto a few weeks ago. As I was standing outside of a Starbucks, I noticed a white BMW stop at the side of the road. The driver stepped out, and at that moment noticed a homeless man sleeping on the side walk. It was extremely cold that day. I was freezing, and I had a sweater and a winter jacket on. The driver of the BMW walked up to the homeless man who was sleeping, took his jacket off, layed it on top of the man, and left. It definitely was unexpected and so encouraging to see such kindness in action. The driver didn't even know I was watching.* Anonymous post.[16]

The previous story shares with the first most of the main characteristics but it differs, for our purposes, in an important one:[17] the probability of role assignment at the beginning of each round. In the first story, in fact, it seems plausible to assume that the two strangers are *ex-ante identical* in their chances of finding themselves without cash in their wallet or, conversely, in their chances to have an extra five pounds —each time they travel by train. On the other hand, in the second story, the BMW driver and the homeless man appear to face, *systematically*, very different chances of being in need of a jacket —each time there is a cold night in Toronto. It appears that, across different cold nights, the BMW driver will have a very high probability of being in the position of the 'helper', while the homeless person (unfortunately...) will have a very high chance of needing a jacket in the role of the 'receiver'. The BMW driver and the homeless man are

[16] Retrieved the 30/7/2012 from http://www.randomactsofkindness.org/kindness-stories/516-jacket-from-the-rich.

[17] The stage game is still a dictator game played among strangers randomly paired from some population.

*heterogeneous* in their probability of needing help and, to the other extreme, of being actually able to give some help.[18]

We accommodate this idea of *persistence* of one's role across different stage games in two distinct ways. First, we allow for the possibility of having *Markov roles*: a helper and a receiver today face different probability distributions over tomorrow's roles (e.g., if I have a jacket to give away today, I will —more likely— have a jacket to give away also tomorrow). Second, we allow for the possibility that different people have different *permanent types*: irrespectively of their role today, different people face different probability distributions over tomorrow's roles (e.g., because of different levels of —say— ability, some people are *always* more likely to be in the role of the helper or in that of the receiver). We show that, within this class of persistence, acts of kindness are more likely to arise in communities characterized by Markov roles than by permanent types. Moreover, with respect to the homogeneous case, both forms of heterogeneity make the social norm of kindness *harder* to sustain at the society-wide level. This theoretical prediction is in line with the empirical evidence: for instance, it has been shown that more heterogeneous communities are characterized by a lower level of social activities (Alesina & La Ferrara, 2000), by a lower level of social trust (Alesina & La Ferrara, 2002), and by a lower provision of public goods (Miguel & Gugerty, 2005). On the other hand, even though stark heterogeneity might impede kindness from thriving at the society-*wide* level, we pursue the possibility of sustaining it *within* more homogeneous *subgroups* of the whole society. This is a well known issue in the social sciences.

Although relatively recent in economics, 'group formation and indirect reciprocity' is a very old and central topic in anthropology and sociology: as Kolm (2001) puts it, rephrasing Gouldner (1960), indirect reciprocity is the "basic glue

[18] We do not want to rule out the possibility that even well off people can find themselves in the position of needing some help and, on the other hand, that people in difficult situations could still be able to help others. Indeed, on the web it is plentiful of stories of this sort. See, for instance: http://www.randomactsofkindness.org/kindness-stories/512-what-is-twenty-dollars, the story of a homeless woman helping out another homeless person; or http://www.helpothers.org/story.php?sid=30619, the story of a homeless woman purchasing a coffee for a wealthy person. (Both stories were retrieved the 30/7/2012.)

that makes people constitute groups or societies." In the same spirit, Mauss (1924) calls reciprocity "one of the human rocks on which societies are built."

Within this body of literature, most of the early economic studies try to assess the groups' ability of sharing *idiosyncratic* risks (i.e., illness, unemployment, poor agricultural performance, etc.) —through indirect reciprocity— among their members.[19] A common finding is that risk-pooling groups appear to be too small in size in order to guarantee full insurance to the participants: typically, these groups are smaller than the efficient society-wide network.[20] This evidence has motivated further empirical and experimental research: *why* and *how* do people form risk-sharing subgroups of the entire society?[21] Arcand & Fafchamps (2012) document, in an empirical study in Burkina Faso and Senegal, *positive* assortative matching on the base of land ownership, education, age and ties with society authorities.[22]

Taking these empirical facts into consideration, we extend our basic framework to highlight two main channels which can prevent the social norm of kindness from spreading at the society-wide level and confine it to smaller groups: limited information about the actual level of kindness in society (which gives rise to free-riding) and, as already mentioned, role's persistence in the form of permanent types.

The structure of the paper is as follows. Section 1.1 clarifies why we consider self-interested agents rather than assuming altruistic or reciprocal preferences and section 1.2 gives and overview of the related literature. Section 2 introduces the baseline model with perfect monitoring and homogeneous agents. The assumption of perfect monitoring is relaxed in section 3 where we allow for the possibility of

[19] See, for example: Townsend (1994), Udry (1994), Jalan & Ravallion (1999), Gertler & Gruber (2002), Murgai et al. (2002), and Fafchamps & Lund (2003).

[20] As long as individuals are risk-averse, shocks are at least partly idiosyncratic and the formation/maintenance of groups is costless, then efficient risk-sharing requires that the risk-pooling group be as large as the economy itself. See, for instance, Fafchamps (2008).

[21] See, for instance: Fafchamps & Gubert (2007), Abramitzky (2008), Barr & Genicot (2008), Fafchamps (2008), Angelucci et al. (2009), Ligon & Schechter (2011), Arcand & Fafchamps (2012), and Attanasio et al. (2012).

[22] Similar people, along each dimension, tend to group together. As an example, assume in the society there are two groups. Consider the case of education. Then, the results suggest that in one group we should observe most of the better educated people and in the other most of those with a lower level of education.

free riding. An agent can refuse to act kindly and go undetected, as long as there remains a sufficient number of agents acting kindly. Chapter 3 considers a different variation of the basic model. Here we introduce persistent differences in risk and the consequences for the sustainability of kindness. In section 1 we consider the case of Markov roles, where agents have an increased likelihood to keep their role from one period to the next. on the other hand, in section 2 we study the case of permanent types, where agents have different probabilities to be a helper or receiver in every period. We find that a society-wide norm of kindness might not be feasible with heterogeneous agents. Consequently, we examine the feasibility and welfare properties of kindness in sub-coalitions. Section 3 concludes both chapter 2 and 3.

1.1. **Why Selfishness?** The fact that we depict instances of kindness such as those in the stories reported above as *in*directly reciprocal outcomes (as opposed to directly reciprocal) follows from restricting our attention to kindness from and to *strangers*. On the other hand, the fact that we represent our players as purely selfish requires some motivation.

In the economic literature, the gift-giving behaviour of people has been motivated by, at the very least, three main classes of preferences: altruistic, selfish, and reciprocal.[23] Broadly speaking, a person has altruistic preferences if she cares both about her own payoff and about the payoffs of others.[24] On the other end of the spectrum, an individual has selfish preferences if she only cares about her own payoff. Differently, a person displays reciprocal preferences if she cares both about her own payoff and about the *behaviour* of other people: in our context, an individual might want to reward those who were seen to help others in the past

[23] For finer classifications of the motives that could possibly drive the gift-giving behaviour of people, see, for instance: impure altruism and warm glow in Andreoni (1989) and in Andreoni (1990); altruism and spitefulness in Levine (1998); kindness (as a motive and *not* as an outcome) and confusion in Andreoni (1995), fairness and inequity aversion in Fehr & Schmidt (1999) and in Bolton & Ockenfels (2000); mimicking in Fowler & Christakis (2010); risk-sharing and consumption-smoothing in Kimball (1988), in Coate & Ravallion (1993), and, for a comprehensive survey, in Fafchamps (2008). For a recent discussion on cooperation from an evolutionary perspective, see Nowak (2012) and the references therein.

[24] This same definition sometimes goes, in the economic literature, under the label of *social preferences*. See, for example: Charness & Rabin (2002).

and to punish those who did not.[25] Given our focus on kindness within pairs of *strangers*, who —by assumption— do not know anything about the *personal* past behaviour of the partner, the reciprocal motive is excluded *a priori*.[26]

It is common in laboratory experiments to observe approximately 50% of the subjects in one-shot dictator games to give away some of their money to anonymous receivers (Camerer, 2003). Because the game is not repeated, this finding is at odds with the predicted behaviour of selfish agents.[27] As Hammond (1975) puts it, it seems evident that altruism is a *sufficient* condition for any charitable behaviour we may observe, but —more interestingly—, is it also *necessary*?

Somehow ignoring Hammond's cautionary question, up until the last fifteen years, these high rates of giving were taken as evidence of the fact that most people are altruistic. In the last fifteen years, experimental economists have investigated more systematically *why* so many subjects decide to give in one-shot dictator games. The evidences gathered so far greatly deflate the altruistic motive: *some* people seem to be truly altruistic, but many more appear to be acting strategically for some expected personal return rather than to make their peers better off. Selten & Ockenfels (1998) is one of the first experiments where the altruistic motive is explicitly challenged, the authors favour a risk-sharing kind of story. Cherry et al. (2002) find that when people have the opportunity to give away *earned* wealth (as opposed to *windfall* wealth), the rate of giving decreases from 80% to 21%. In addition, when people are also given the chance to free-ride, the giving rate falls further to 3%. Dana et al. (2007) confirm that adding the possibility of free-riding approximately halved the giving rate. List (2007) and Bardsley (2008) show that by enlarging people's action set to *taking* money (on top of giving) nearly all giving vanishes. Ligon & Schechter (2011) show that many people, in a field experiment with windfall money, are observed to give gifts out of altruism

[25] For discussions on reciprocal preferences/motives see, for example: Rabin (1993), Nowak & Sigmund (1998), Charness & Rabin (2002), and Dufwenberg & Kirchsteiger (2004).

[26] As explained in greater detail in the literature review, we distinguish between reciprocal *outcomes* and reciprocal *preferences/motives*. In principle, reciprocal preferences are a sufficient but *not* a necessary condition for reciprocal outcomes. As a consequence, by excluding *a priori* the reciprocal motive, we are *not* ruling out the possibility of obtaining reciprocal outcomes.

[27] For a discussion see, for example: List (2007) and Bardsley (2008).

but that, in a more realistic environment (with earned money), that evidence disappears. We interpret these results as suggestive of the fact that there exist conspicuous shares of gift givers who are not, in our terms, genuinely altruistic or, more precisely, whose altruistic motive —compared to the selfish— is not a particularly robust/consistent. Still, the aforementioned evidence only partially addresses Hammond's question: it confirms that in one-shot dictator games we should not observe such high rates of giving, but it does not show that selfish people would indeed engage in charitable behaviour. To do that, we turn to experiments in *repeated* dictator games.

Seinen & Schram (2006) is the first paper which experimentally investigates the environment we model: a repeated dictator game with, at the beginning of each round, random matching in pairs from a large population and random assignment of roles (helper and receiver) within each pair. In a treatment group, people know the history of play of the agent with whom they are matched (public histories); in another treatment group —as in our framework— this information is not given (private histories), i.e., the two paired players are *strangers*. The researchers find evidences of indirectly reciprocal outcomes among strangers: the giving rate is 18%. Also, people do react to external incentives: when histories are public, the helping rate grows to 74%. Englemann & Fishbacher (2009) perform a complementary experiment in which the aim is to further investigate the result on the public histories treatment: are people helping more out of reciprocity or out of selfishness?[28] For this end they include the possibility that only half of the agents in the population have public histories. The experimenters find confirming evidences of indirectly reciprocal outcomes among strangers: the giving rate is 32%. Furthermore, they obtain that 80% of subjects react to strategic incentives: they document an increase of 5 points in the helping rate due to reciprocity and of 35

[28] In this context, the authors define a selfish person as someone who helps in order to positively affect her *own* history of play with the aim of increasing her likelihood to be helped, if needed, in the future (i.e., strategic reputation-building). This *irrespectively* of the history of play of the receiver with whom they are paired.

points due to selfishness. Charness & Genicot (2009), in a context of directly reciprocal outcomes,[29] find evidences that favour the selfish over the altruistic motive: people exchange gifts in order to smooth consumption when facing uncertainty, relationships which are expected to last longer give rise to higher rates of giving, as do higher degrees of risk aversion. Leider et al. (2009), in a large field experiment conducted at Harvard dormitories, document the relevance of indirectly reciprocal outcomes among strangers and that, similarly to Charness & Genicot (2009), the expectation of a longer/more intense relationship positively affects the present rate of giving.

In the last years, economists have started investigating how gift-giving behaviour is affected by social networks. Specifically, there has been a growing interest in testing the intuitive hypothesis that altruism towards a person gets stronger the tighter the relationship with this person (*directed* altruism). Hoffman et al. (1996) suggest that a decrease in perceived social distance increases the giving rate in dictator games. Leider et al. (2009) find that, when helpers are paired with receivers who are close friends (as opposed to strangers), their gift-giving rate motivated by altruism increases by 52%. Ligon & Schechter (2011) find qualitatively similar results. Directed altruism is also proposed by Fafchamps & Lund (2003) as an interpretation for their empirical findings. Along the same lines are the results from the field experiment conducted by Attanasio et al. (2012). Moreover, Fafchamps (2008) collects empirical evidences that depict altruism as being limited to relationships between close relatives.

Returning to Hammond's question, we read the experimental evidences gathered so far as an indication that altruism is, even though a sufficient, *not* a necessary condition for charitable behaviour. On the one hand, many people show to have, at some level, altruistic preferences; but on the other, many of these seem to respond readily to incentives that affect their own payoffs in ways that imply a stronger and more stable selfish motive (over the altruistic). In addition, the fact

[29] The *same* pair of agents play repeatedly a dictator game with random assignment of roles at the beginning of each round.

that altruism seems to be *directed* further diminishes its appeal in our specific context: a population of strangers.

1.2. **Related Literature.** The current article is related to different streams of economic literature. In what follows, we briefly review each of them and highlight the main differences with respect to what we do.[30]

The single article that inspired our baseline model the most is Hammond (1975). There, he proposes a repeated dictator game (called "poverty game") in which directly reciprocal outcomes (i.e., only two agents) between selfish agents can be sustained in equilibrium due to the fact that agents switch roles (i.e., helper and receiver) at each round. In our baseline model, we extend Hammond's idea to indirectly reciprocal outcomes (i.e., more than two agents) and generalize the role switching from deterministic (i.e., each round agents switch roles for sure) to stochastic, with and without persistence (i.e., each round agents have some probability of switching roles).[31]

A first relevant literature is that on 'random matching games', started by Rosenthal (1979) and Rosenthal & Landau (1979). Kandori (1992) and Okuno-Fujiwara & Postlewaite (1995) greatly enrich the initial findings and, importantly, extend the Folk Theorem for this class of games. Kandori (1992) draws attention to a key relationship between information and cooperation: if information circulation is somehow limited within society, then the larger the size of society, the harder it gets to sustain cooperation. Ellison (1994) further develops upon this point. Since then, a constantly growing body of research has been investigating the issue: what are the institutions that can help the survival of cooperation in large communities with little or no information circulation about agents' past behaviour? One of the first efforts in this direction is Gosh & Ray (1996). The authors show that if pairs are not re-matched each round at random (i.e., agents are free to build up long-lasting relationships) and there is some heterogeneity in the population, then

[30] As in our framework, a common assumption shared by all the papers cited in this section is that contracts are not enforceable. As a consequence, only self-enforcing outcomes can be sustained in equilibrium: this requires that at any point in time, the benefit from complying with an agreement must outweigh the gain from reneging.

[31] We study Hammond's role switching as a special case of the model with Markov roles.

cooperation can be sustained in large communities of strangers. Watson (1999) is another notable example of cooperation sustained by a mix of reputation-building and heterogeneity. More recently, Athey et al. (2010) propose various mechanisms where cooperation is achieved through reputation-building alone (i.e., no agents' heterogeneity): group-specific investments and social hierarchies. The model allows agents to choose, prior to the random matching phase, the group within which the stage game will be played. In other words, the whole society is divided into groups and agents choose, each round, which group to attend. As a consequence, —on the one hand— the authors can exploit agents' group choice histories (i.e., agents' 'seniority' in each group) to engender within-group loyalty and construct cooperative equilibria. On the other, they can address the issue of endogenous group formation. This latter line of research, about informal society division (e.g., castes, tribes, clans, etc.), is further explored —for instance— by Choy (2013). The author extends the framework proposed by Gosh & Ray (1996)[32] adding two dimensions: first, society is partitioned into groups and second, agents can observe if their partners have ever interacted with members of other groups. Then, he shows that the social norm prohibiting agents from forming relationships with members of different groups (i.e., group segregation) can be welfare improving and thus, able to persist.

The main theme of our paper is different from that of the random matching literature, even though —from a modelling perspective— the commonalities are twofold. Our main interest does not lie in studying the minimal information transmission mechanisms necessary to sustain efficient outcomes by society enforcement. Rather, we investigate how risk about agents' future roles or positions in society shapes their current attitude towards cooperation. On the other hand, our models do share common features with those in the random matching literature. First, in an extension of our baseline model, we limit the circulation of information about past levels of kindness in society —à la Green & Porter (1984)—[33] and

[32] Consequently, cooperation is obtained —as in the original model— through a mix of reputation-building and agents' heterogeneity.

[33] The influential model proposed by Green & Porter (1984) can be summarized, for our purposes, as follows. A group of firms produce a homogeneous product and compete on quantities. The

operationalise, in our somehow different framework,[34] the basic idea suggested by Kandori (1992): the survival of the social norm of kindness in large communities can be endangered by free riding. Second, we suggest an alternative trade-off (with respect to Kandori's): between heterogeneity and cooperation. The more heterogeneous is society, the harder it gets to sustain cooperation. In studying this relationship, we assume perfect circulation of information; consequently, we do not implement any reputation-building mechanism in order to sustain cooperation.[35] Even though society-wide cooperation might not be possible because of stark heterogeneity, less diverse subgroups of agents could still cooperate.

A second relevant literature is that on 'informal risk-sharing' started by Kimball (1988) and Coate & Ravallion (1993). These models depict small communities in which agents are subject to volatile streams of non-storable income and, in order to obtain some degree of insurance, agents organize a 'common pot' of money: in each period, all those with 'high' income deposit money in the 'pot', while those with 'low' income collect money from the 'pot'. In these communities agents are homogeneous, information circulation is perfect, and income-sharing agreements are not enforceable (hence informal). Because of the lack of commitment, informal contracts regulating the functioning of the 'common pot' must be self-enforcing.

oligopolists sell their product at a common price which is a function both of the quantities injected in the market by all the competitors and of an independent stochastic element (i.e., some price shock determined independently of firms' behaviours). Firms observe the aggregate price but not its individual components. As a consequence, when the product's price is 'low', the oligopolists do not know with certainty if that is a consequence of someone selling 'a lot' (i.e., above the Cournot level) or, conversely, of some price shock. In this environment, in order to sustain a profitable collusion in equilibrium, the authors construct a grim-trigger strategy where firms have to infer from the observed price and their knowledge of the stochastic element's distribution how plausible it is that no one is 'misbehaving' by producing more than their due quantity. Our free riding extension is inspired by this mechanism.

[34] Most of the models in the random matching literature, in the post-Kandori (1992) era, employ a prisoner's dilemma as a stage game. Conversely, we use a dictator game with random role assignment.

[35] Rohner (2011) proposes a model of social tensions in which a similar channel is at work: social disputes are particularly likely to arise in populations that are more ethnically heterogeneous. Furthermore, the author shows —in his different framework— a result close in flavour to one of ours: when the dividing line in society is class rather than ethnicity, where contrary to the latter the former is not 'immutable', fewer social tensions are predicted. Parallely, in our model with ex-ante heterogeneity: permanent types jeopardize society-wide kindness more than Markov roles. Rohner's (2011) model differs from ours in many respects, one above the others: cooperation hinges on reputation-building.

The possible definitions of 'self-enforcement' and the corresponding degrees of insurance they give rise[36] represent the object of this literature. The early papers, such as Coate & Ravallion (1993), Kocherlakota (1996), Kletzer & Wright (2000), and Ligon et al. (2002), focus on concepts of self-enforcement which are mainly ex-post stable: they consider as relevant only individual deviations from society-wide cooperation, ignoring the possibility that subgroups of the entire society could break the informal society-wide agreement and decide to create their own common pot. Stemming from this observation, Genicot & Ray (2003) tighten the definition of self-enforcing contract by adding a requirement of ex-ante stability: a common pot can be created among a group of agents only if there is no subgroup which would unilaterally profit from creating its own, smaller, common pot. Bold (2009) builds upon the framework of Genicot & Ray (2003) and fully characterizes the set of coalition-proof informal agreements. Bloch et al. (2007) propose a complementary solution concept with respect to the existing ones, that of fragility: they allow for deviations to occur in some states and then pose bounds on the probability that a deviation is observed. Another example of endogenous group formation within this class of models is Weynants (2011). The most recent evolution in the informal risk-sharing literature is represented by Bloch et al. (2008) and Ambrus et al. (2010). The researchers move away from the notion of insurance-group toward that of insurance-network. They do so because recent empirical evidence suggests that a significant segment of (informal) insurance transactions is bilateral (i.e., directly reciprocal outcomes).[37]

Our paper is conceptually related to the informal risk-sharing literature. This is so because of the viewpoint on kindness we propose: a non-enforceable and indirectly reciprocal outcome sustained by the selfish motive of intertemporal consumption-smoothing. There is one main difference between our baseline model

[36] From perfect-sharing, in which everyone contributes to the common pot with everything, to autarky, in which each agent is left with her own income.

[37] In our case, because agents are randomly re-matched in each period, we believe that the notion of group is more appropriate then that of network.

and that employed in this literature: the random matching element.[38] In the informal risk-sharing model, within each period, agents endowed with 'surplus' money can transfer it to anyone with a 'deficit' anywhere within society, costlessly.[39] Thus, the possibility of helping is not pair-specific.[40] Differently, in our framework this is not the case: acts of kindness are assumed to be non-transferable and, consequently, pair-specific. An agent in need can only be helped by the agent she bumps into if the latter actually is in the position to give some help, irrespectively of the fact that there could be many other potential helpers around the corner. Conversely, a potential helper can only exert an act of kindness if she bumps into an agent in need, no matter how many agents in need could be waiting elsewhere in society.[41] [42]

A third relevant literature is that on 'enforced reciprocity'.[43] Two notable examples are Karlan et al. (2009) and Leider et al. (2009). These articles represent a mired attempt to conceptually and empirically disentangle the reciprocal preferences/motives from the reciprocal outcomes. Referring, in general, to 'reciprocity'

[38] To be precise, Bloch et al. (2007) do have random matching in their model. On the other hand, the aim of their paper is very distant from ours: they propose —in very general terms— a new stability concept, that of fragility.

[39] The same is not true across time periods because income is assumed to be non-storable. Similarly, in our framework "There's no use doing a kindness if you do it a day too late" (to express the idea with the words of Charles Kingsley).

[40] Notice that, here, we are abstracting from the incentives of the agent who is endowed with the extra income to be willing to share her sums with the rest of society. Our argument is a priori with respect to any consideration of willingness, it has merely to do with the action set available to the members of society: independently of her fondness, an agent with 'high' income could always (and costlessly in this model) contribute with part of her endowment to the common pot. Equivalently, any 'low' income agent could always (and costlessly) collect money from the common pot.

[41] Again, this is a priori with respect to any incentive-related consideration.

[42] Clearly, ours is a subjective modelling choice. We decide to concentrate on unplanned and spontaneous acts of kindness between pairs of strangers. Nothing prevents agents from organizing something similar to a 'common pot' also in the context of acts of kindness. We actually observe many of these initiatives in the form of charitable organizations. For instance, a group of people could systematically collect coats from wealthier individuals and, during the coldest nights, purposedly go around the city and distribute them to homeless people in need. Whenever an act of kindness is somehow planned, intermediated, and stimulated by a third party (say, by a charitable organization), then the economic circumstances within which a potential helper must decide her action rapidly increase in complexity with respect to our stylized framework (see, for example: social pressure in Della Vigna et al. (2012) and framing in Grossman & Eckel, 2012). In this case, we believe that —from a modeling perspective at least—, the informal risk-sharing framework is closer to the target and the interested reader is forwarded to the aforementioned literature.

[43] Additional discussions on reciprocity (with related references) can be found both in the introduction and in the section "Why selfishness?"

can be a source of ambiguity. We could be talking either about reciprocal outcomes or about reciprocal motives. These are two very different objects. For instance, in our context: reciprocal outcomes are the observable acts of kindness among the members of a group, while reciprocal motives are the unobservable preferences that could —potentially but not necessarily— motivate the reciprocal outcomes. Karlan et al. (2009) propose a tractable model where reciprocal outcomes can be obtained without requiring agents to have reciprocal preferences. Leider et al. (2009) find that the model fits their experimental data better than alternative preference-based theories of reciprocity. Even though both the papers are quite distant from ours, we still believe to share with them some perspective: we interpret acts of kindness as indirectly reciprocal outcomes obtained through selfish motives.[44]

## 2. Basic Model

In the current section we lay out the baseline model: a formalization of the first story reported in the introduction. Three main assumptions underline the baseline model: perfectly anonymous information, absence of persistence in agents' roles, and population homogeneity. After studying the baseline model, we will relax the assumptions one by one in the following sections of the article.

2.1. **Description of the Game.** Within a society of strangers,[45] dictator games are played for infinitely many periods. Agents are randomly matched in pairs at the beginning of each period. Every agent, in each period, can be of two possible roles: a helper (H) or a receiver (R). At the beginning of each period, Nature allocates a role to every agent in society following the objective distribution (known to the agents): $\Pr(\mathrm{R}) = p$ and $\Pr(\mathrm{H}) = 1 - p$. This probability distribution is exogenously given (i.e., agents cannot do anything to affect the likelihood of having a specific role) and independent of agents' previous roles (no persistent roles).

[44] For an in-depth discussion on "Why selfishness?" (as opposed to altruism or to reciprocity, for example), see section 1.1.

[45] For the time being, a precise specification of society is not essential. We could either have, for instance, a continuum of agents or a finite number of them. In the later extensions of the basic framework, whenever required, we will be more precise.

Note that $p$ can also be interpreted as the expected proportion of receivers in society, and thus the expected likelihood to be matched with a receiver. Each agent observes both her own and her partner's role realization. The combination of roles in a pairing determines the agents' action sets and payoff functions. Define any (H,R) or (R,H) match as a relevant match. Only in relevant matches is a dictator game played.

(1) Relevant Match: (H,R) or (R,H)
  - H can either play kind (K) or not kind (NK). R has an empty action set.
  - If H plays K, then H incurs the cost $c$ and R receives the benefit $b$, where $b, c > 0$.
  - If H plays NK, then both players get a payoff of zero.

(2) Other Matches: (H,H) or (R,R)
  - No game is played, both players have an empty action set.
  - Both players get a payoff of zero.

Without loss of generality, $c$ is normalized to 1. The agents know if all helpers in relevant matches have played K in all past periods or if someone has ever played NK. However, this information is not personal, agents do not know how a specific player behaved in the past. Moreover, they do not keep any record of the identity of the people with whom they played dictator games in the past. In other words, whenever two agents meet more than once, they will not recognize each other and think they are in the presence of a stranger (perfectly anonymous information).

2.2. **Equilibrium.** Notice that if the game were not repeated (so the agents only played once), then the unique Nash Equilibrium (NE) would be for the helper in a relevant match to play NK. Also, if the game were infinitely repeated and Nature moved only at the beginning of the first period (so if each agent kept the same role along the whole game), then the unique Subgame Perfect Nash Equilibrium (SPNE) would be for a helper to play NK every time she is in a relevant match.

Differently, in the game with Nature moving at the beginning of each period cooperative outcomes can be sustainable in equilibrium.[46] Specifically, the following strategy can be an SPNE of the game:

> *If you are the helper in a relevant match, play K unless at least one helper in a relevant match played NK in the past. Otherwise, if at least one helper in a relevant match played NK in the past, play NK.*

We define the cooperative equilibrium outcome as the social norm of kindness.[47] Note that, if all agents follow the proposed strategy, a deviation by a single agent leads to the breakdown of the social norm of kindness forever. Thus, this amounts to a collective grim trigger strategy. The only agents who have the possibility to deviate are, in any given period, helpers matched with receivers. Given that everyone else has played K so far, deviation (NK) is not profitable for a helper in a relevant match if:

$$-1 + \sum_{t=1}^{\infty} \delta^t \left[ p \cdot (1-p) \cdot b - (1-p) \cdot p \right] \geq 0,$$

where $\delta \in (0,1)$ is the discount factor common to all agents. Hence, the cost of being kind needs to be smaller than the expected future payoff from sustaining the social norm of kindness. This can be rewritten as:

$$b \geq 1 + \frac{1-\delta}{\delta \cdot p \cdot (1-p)} = b^{Base}(\delta, p) \ . \tag{8}$$

[46] At the same time, the repetition of the stage equilibrium, where agents always play NK, is always an equilibrium of the repeated game. In fact, there can be many equilibria, where agents play NK some of the time. However, we do not consider the issue of equilibrium selection here and focus on the fully cooperative equilibrium.

[47] Notice that we are not referring to any specific strategy in the current definition. Any equilibrium strategy which gives rise, on the equilibrium path, to the cooperative outcome is consistent with the social norm of kindness.

$b^{Base}$ gives the minimum value of $b$ which is necessary to sustain kindness. Notice that $b^{Base}$ is decreasing in $\delta$ and as $p$ approaching $0,5$. In other words, $b$ can be lower if agents are more patient and if the likelihood of becoming a receiver is neither too low nor too high. Indeed, if an agent is patient, she is more willing to be kind today in the expectation of receiving kindness in the future. Also, in order for a helper to be willing to exercise kindness today, she must believe that in the future both the possibility of being in need of kindness is concrete and that, in this case, there is a sufficient likelihood of meeting a helper.[48]

## 3. Imperfect Monitoring

In this section, we relax the assumption of perfect monitoring maintained in the basic model. Differently from the baseline model, agents do not know if every helper in a relevant match actually played K in the previous period. They observe the aggregate number of K played in the previous period but do not know what the number of relevant matches was. A practical example for this could be that there is information about the aggregate amount of charitable donations in a given year and agents make inferences about the level of charity in their community. There will be an equilibrium where the social norm of kindness can be sustained, as long as the number of acts of kindness stays above a critical threshold. When the level of kindness falls below the threshold cooperation breaks down. When considering whether to cooperate and behave kindly as a helper in a relevant match, an agent considers how likely it is that she is pivotal for the sustainment of the norm. Only if a failure to be kind leads to the breakdown of the kindness equilibrium, does the agent feel a negative consequence of her deviation. With respect to the basic

[48] In the attempt of keeping the algebra as simple as possible while preserving the economic message of the model, we derive the condition for existence of a cooperative SPNE only for the grim trigger strategy. This gives us the lowest threshold for $b$. If $b > b^{Base}$, additional cooperative equilibria with a finite number of punishment periods exist. However, on the equilibrium path of the models with perfect anonymous information, there will be no deviation and thus no need for punishment. Thus, different punishment lengths are observationally equivalent and imply the same level of welfare. Welfare is only affected in the free riding extension, where cooperation can break down on the equilibrium path.

model, this introduces the possibility of free riding and, consequently, a new trade-off. We can interpret the social norm of kindness analogous to a public good. The incentive to play K comes from the requirement to reach the threshold to sustain the social norm for the future when the agent could be in need of kindness herself. A free rider would be helper in a relevant match who plays NK without causing the social norm of kindness to break down (i.e., someone who does not contribute to the public good when asked, but who still benefit from it when in need).[49] It will be shown that this kind of imperfect monitoring can lead to a limit on the size of the population where social norm of kindness can be sustained. As the size of society gets larger, it becomes harder to sustain the norm, since it is less likely that an agent is pivotal (CLAIM 2 below). On the other hand, it is precisely when society size grows that the social norm of kindness becomes —potentially— more gainful (CLAIM 5 below).

3.1. **Description of the Game.** There is a society of $2 \cdot N$ agents divided into two independent groups of equal size, $N$. At the beginning of each period, each of the $N$ agents in one group is randomly matched with another agent from the other group, so there are $N$ matches per period. As in the basic model, each agent is allocated a role [either helper (H) or receiver (R)] at the beginning of each period, with$\Pr(\text{R}) = p$ and $\Pr(\text{H}) = 1-p$. For each agent, this probability is exogenously given and independent of her previous period's role (no persistent roles). As in the baseline model, the case when helper and a receiver meet is defined as a relevant match. Dictator games are played only in relevant matches. The probability that a given match is relevant is $q = 2 \cdot p \cdot (1-p)$. Then the number of relevant matches out of $N$ in a given time period is distributed according to a binomial distribution with sample size $N$ and success probability $q$. Define the number acts of kindness (number of agents playing K) in a given time period as $K_N$. If all helpers in a relevant match cooperate, $K_N$ obviously has the same distribution as

[49] An alternative possibility to introduce imperfect monitoring would be that a failure to be kind would be observed and become known with a certain probability. Also in this case it would make sense that detection becomes less likely as group size increases. Thus our setting is only one of many possibilities to capture these issues. The information structure bears a resemblance to Green and Porter (1984), where companies infer from price changes how likely it is that another company has deviated from a cartel.

the number of relevant matches: $K_N \sim Binomial\,(q, N)$ with $\mathbb{E}\,[K_N] = N \cdot q$ and $\mathbb{V}ar\,[K_N] = N \cdot q \cdot (1 - q)$. The PMF of $K_N$ is given by $f_N$ and the CDF as $F_N$. The action sets and the payoff functions are the same as in the basic game.

Agents can make inference from the realization of the random variable $K_N$ about the attitude towards kindness of society as a whole. They know the probability distribution of the number of relevant matches and observe their own role as well as the role of the agent they meet in each period. However, they do not observe the realized number of relevant matches in each period. Furthermore, agents observe the number of K's that were played in the previous period, $K$ —which is a realization of the random variable $K_N$. We think it is realistic to assume that agents receive some signal about the overall 'level' of kindness in society beyond their narrow personal experience, and that this affects their own attitude towards kindness.

We assume that agents do not carry over time the memory of their personal histories of play: the only information helpers in relevant matches process while deciding how to play is last period's realization of $K_N$. This assumption is a simplification that - if anything - will make kindness harder to sustain. Agents with memory could only face a further incentive to play K, therefore the absence of memory is both analytically convenient and enables us to isolate the free riding incentive from confounding elements.[50]

3.2. **Equilibrium.** We consider SPNE of the following form: All agents have the same critical number number of acts of kindness $k^\star$ and follow the strategy:

[50] Consider e.g. a strategy where agents seize cooperation both after K falls below the critical level as well as after personally experiencing defection in one of their personal interactions. Given that everybody else seizes cooperation after the critical level has been reached, it is always subgame perfect to do so as well. If agents also stop cooperation after personally experiencing defection, this would constitute an additional channel through which defection can lead to negative consequences for the perpetrator. However, - as shown by Kandori (1992) - it might not be subgame perfect to stop cooperation after personally experiencing defection for some parameter values. In this case cooperation only breaks down if K falls below the critical value and the incentives are exactly the same as in the setting without memory. Therefore, memory would only make it easier to sustain cooperation.

> *If you are the helper in a relevant match, play K if minimum was exceeded ($k \geq k^{\star}$) in the previous period. Otherwise ($k < k^{\star}$ in the previous period) play NK.*

Thus the game can be in two possible phases. In the cooperative phase the social norm of kindness is intact. The minimum level of kindness $k^{\star}$ has been reached in all past periods and thus all helpers exercise kindness when matched with a receiver. In the uncooperative phase the social norm of kindness has broken down. In one previous period the minimum level of kindness $k^{\star}$ has been missed, so no one exercises kindness any more. Clearly, the suggested strategy is subgame perfect in the uncooperative phase. Next, we check for which parameters it also holds for the cooperative phase. For conciseness, define:

$$\theta = p \cdot (1 - p) \cdot (b - 1) \, ,$$

where $\theta$ gives the expected per-period-payoff in the cooperative phase. Further, define the ex-ante probability of continuing the cooperative phase for one more period, provided that all agents follow the equilibrium strategy:

$$\alpha = \Pr \left( K_N \geq k^{\star} \right) = 1 - F_N \left( k^{\star} - 1 \right) .$$

Then the continuation value for of playing K for a helper in a relevant match is:

$$V_K \;\; = -1 + \beta \cdot \sum_{t=1}^{\infty} \left( \delta^t \cdot \alpha^{(t-1)} \cdot \theta \right) .$$

The agent has to pay a cost of 1 and has and expected payoff of $\beta \cdot \sum_{t=1}^{\infty} \left( \delta^t \cdot \alpha^{(t-1)} \cdot \theta \right)$ from continuing the cooperative phase in future periods. Here, $\beta$ stands for the probability of continuing the cooperative phase to the next period, given that the concerned agent plays K and provided that all other agents follow the equilibrium strategy. Then, $\beta$ is given by:

$$\beta = \Pr\left(K_{N-1} \geq k^\star - 1\right) = 1 - F_{N-1}\left(k^\star - 2\right),$$

where $F_{N-1}$ is the CDF of $\mathit{Binomial}\left(q, N-1\right)$. This is because, if the respective agent plays K, there only have to be a minimum of $k^\star - 1$ acts of kindness in the remaining $N-1$ matches in order to reach the critical value $k^\star$. After the next period, the probability of continuing the cooperative phase from one period to the next is given by $\alpha$.

On the other hand, if the agent defects and plays NK, the continuation value is:

$$V_{NK} = 0 + \gamma \cdot \sum_{t=1}^{\infty} \left( \delta^t \cdot \alpha_N^{(t-1)} \cdot \theta \right) .$$

In contrast to $V_K$, the agent does not pay the cost of kindness and the probability of continuing the cooperative phase to the next period is now given by $\gamma$. This probability will be lower, given that the agent now plays NK. $\gamma$ is then given by:

$$\gamma = \Pr\left(K_{N-1} \geq k^\star\right) = 1 - F_{N-1}\left(k^\star - 1\right),$$

This is because, if the respective agent *does not* play K, there now have to be a minimum of $k^\star$ acts of kindness in the remaining $N-1$ matches in order to reach the critical value $k^\star$. After the next period, the probability of continuing the cooperative phase from one period to the next is again given by $\alpha$. Thus, there will only be a punishment for defection if this leads to a breakdown of cooperation. Note that the difference in continuation probabilities is:

$$\beta - \gamma = F_{N-1}(k^\star - 1) - F_{N-1}(k^\star - 2) = f_{N-1}(k^\star - 1),$$

which is just $\Pr(K_{N-1} = k^\star - 1)$, the probability that there are exactly $k^\star - 1$ acts of kindness in the remaining $N-1$ matches. If this is the case, then the agent is pivotal. The cooperative phase continues only if she plays K and breaks down if she plays NK. Only in this case there will be a negative consequence for the deviator.

Hence, in the cooperative phase, a deviation (NK) is not profitable for a helper in a relevant match if:

$$\begin{aligned} b &\geq 1 + \frac{1 - \delta \cdot \alpha}{\delta \cdot \frac{q}{2} \cdot (\beta - \gamma)} \\ &= 1 + \frac{1 - \delta \cdot [1 - F_N(k^\star - 1)]}{\delta \cdot \frac{q}{2} \cdot f_{N-1}(k^\star - 1)} \qquad (9) \\ &= 1 + \frac{1 - \delta \cdot [1 - F_N(k^\star - 1)]}{\delta \cdot \frac{k^\star}{2 \cdot N} \cdot f_N(k^\star)} = b^{Free}(\delta, N, q, k^\star). \end{aligned}$$

Equation 9 gives the critical value $b^{Free}(\delta, N, q, k^\star)$.[51] If the benefit of kindness $b$ exceeds this value, then the proposed strategy is a SPNE of the game. Note that $b^{Free}$ depends on the selected minimum value $k^\star$. There can be various possible values of $k^\star$ which can support an equilibrium for a given level of $b$. The selection of the optimal minimum value $k^\star$ is the topic of section 4. In the remaining part of this section we study the behaviour of $b^{Free}$ with respect to changes in the values of the parameters. Proofs for the following claims are reported in Appendix B.

*Claim* 1. As $\delta \to 0$, $b^{Free}(\delta, N, q, k^\star) \to +\infty$.

Not surprisingly, as agents get more impatient, it becomes less attractive to incur the cost of kindness in order to sustain the social norm of kindness for a potential benefit in the future. Thus a higher benefit $b$ is necessary to sustain the equilibrium.

*Claim* 2. As $N \to +\infty$, $b^{Free}(\delta, N, q, k^\star) \to +\infty$.

When society size increases, the probability of being pivotal for the continuation of the social norm of kindness, $f_{N-1}(k^\star - 1)$, gets smaller and a higher benefit $b$ is necessary to sustain the equilibrium. The social norm of kindness can be sustained more easily in smaller communities.

*Claim* 3. $\dfrac{\partial b^{Free}(\delta, N, q, k^\star)}{\partial q} < 0$.

Notice that the maximum value of $q$ is 0.5, when $p = 0,5$. Thus, as the probability of having the role of a receiver and being in need of kindness moves towards 0.5, the social norm of kindness can be sustained more easily. The intuition is the same as that in the basic model: If $p$ is too small, the likelihood of being in need

[51] The third inequality follows from $\begin{pmatrix} N-1 \\ k^\star - 1 \end{pmatrix} = \frac{k^\star}{N}\begin{pmatrix} N \\ k^\star \end{pmatrix}$, given the pdf of the binomial distribution $f_N(k) = \begin{pmatrix} N \\ k \end{pmatrix} q^k (1-q)^{n-k}$.

of kindness in the future is small. If $p$ is too large, the likelihood of meeting a helper when in need is small. In both cases, the preservation of the social norm of kindness is not very valuable.

## 4. Welfare

As noted before, if the benefit level $b$ is sufficiently high, it can support an equilibrium with various possible values of $k^\star$, the critical value of acts of kindness for which the norm is sustained. We define the set $\mathcal{K}$ as the set of critical values $k^\star$ for which the proposed strategy is a SPNE. The purpose of this section is to determine the optimal value $k^\star$, which maximises expected welfare. As in the basic model, also here any equilibrium strategy compatible with the social norm of kindness is strictly preferred to the non-cooperative equilibrium from an expected payoff viewpoint. But, differently from the basic model, it is possible for the the social norm of kindness to break down on the equilibrium path, i.e. without any deviation. This is so because the game has an infinite horizon and $\Pr\left(K_N < k^\star\right) > 0$ for $k^\star > 0$. Even if everyone is following the proposed strategy, it is possible that the number of relevant matches is below $k^\star$ and thus the minimum number of acts of kindness is not met. In this case, the social norm of kindness breaks down not because of the misbehaviour of some agents but simply because the realized number of relevant matches is 'too low' given the values of $q$ and $N$.

Notice how, in statistical terms, $\Pr\left(K_N < k^\star\right)$ given $K_N \sim \mathit{Binomial}\left(q, N\right)$ could be interpreted as the probability of type I error in a test procedure.[52] In fact, the proposed strategy works as if every agent, after having collected the data $k$, performed a statistical test to evaluate the maintained hypothesis that the social norm of kindness holds in the society, $H_0 : K_N \sim \mathit{Binomial}\left(q, N\right)$, whose rejection region is $\left\{k \in [0, N] \middle| \, k < k^\star\right\}$. It follows that the probability of type I error is:

$$\Pr\left(K_N < k^\star \middle| H_0\right) = F_N\left(k^\star - 1\right).$$

[52] The type I error in a test procedure happens when the statistician rejects her maintained hypothesis even though it is correct.

Given this interpretation, we can define $F_N(k^\star - 1)$ as the probability of involuntary breakdown of the social norm of kindness.

Whenever the social norm of kindness breaks down, each agent gets a stream of 'zeros' thereafter. The occurrence of this eventuality with positive probability in equilibrium (i.e., typically $k^\star = 0$ does not sustain the social norm of kindness) is the 'price' that has to be paid in order to constrain free riding attitudes in society. Consequently, it becomes apparent how the choice of $k^\star$ plays an active role also in terms of welfare.

The ex-ante expected payoff of any agent is:

$$\begin{aligned} V &= \alpha \cdot (\theta + \delta \cdot V) \\ &= \frac{\alpha \cdot \theta}{1 - \alpha \cdot \delta} \end{aligned} .$$

It follows that the optimal minimum value $k^\star$ is given by:

$$k^\star = \arg\max_{k^\star \in \mathcal{K}} \left\{ \frac{[1 - F_N(k^\star - 1)] \cdot \theta}{1 - \delta \cdot [1 - F_N(k^\star - 1)]} \right\}, \tag{10}$$

where the set $\mathcal{K}$ collects all the values of $k^\star$ for which an equilibrium can be sustained, as $b \geq b^{Free}(\delta, N, q, k^\star)$, from 9.

*Claim* 4. Assume $\mathcal{K}$ is non-empty. Then the unique solution of 10 is the smallest $k^\star \in \mathcal{K}$.

Because $V$ is strictly increasing in $\alpha = 1 - F_N(k^\star - 1)$, the probability to continue the social norm of kindness, conditional on the proposed strategy being an equilibrium (i.e., a non-empty $\mathcal{K}$), the lowest possible $k^\star$ is the optimal minimum value. Clearly, the smallest minimum value $k^{\star}$ is the one for which the probability of an involuntary breakdown of the social norm of kindness, $F_N\left(k^{\star}-1\right)$, is minimal.

*Claim* 5. For any given $k^{\star}$, as $N$ rises, $V$ increases (given that $b$ is sufficiently large to support the equilibrium).

For any $k^{\star}$, as the size of society gets larger, the expected welfare increases. This is so because, for given $k^{\star}$, the probability of involuntary breakdown of the social norm of kindness, $F_N\left(k^{\star}-1\right)$, decreases in $N$ and, as underlined above, $V$ is strictly increasing in $\alpha=1-F_N\left(k^{\star}-1\right)$. Clearly, when a larger group of agents play the proposed game, the likelihood of an insufficient number of good matches to realise decreases.

On the other hand, as seen in CLAIM 2, a larger $N$ makes the proposed strategy harder to be sustained as an equilibrium. This is because $b^{Free}\left(\delta, N, q, k^{\star}\right)$ increases in $N$, so that a higher benefit $b$ is necessary to ensure the existence of the cooperative equilibrium and to discourage free riding. As $N$ grows large, the number of elements in the set $\mathcal{K}$ decreases, and if the lower bound $k^{*}$ rises, then the overall effect on $V$ is ambiguous. Therefore, an increasing $N$ leads to a trade-off between existence and efficiency.

Finally, it should be remarked that welfare could be enhanced by reducing the length of the uncooperative phase. For simplicity, we have restricted the analysis to a grim trigger strategy, where the social norm of kindness breaks down indefinitely after the number of acts of kindness has fallen short of the minimum level $k^{\star}$. As this is possible even on the equilibrium path, expected welfare is reduced. For a given $k^{\star}$ and a sufficiently high benefit $b>b^{Free}\left(\delta, N, q, k^{\star}\right)$, the uncooperative phase could be shortened to a finite number of periods $T$. However, in this setting there would be a trade-off between the minimum level of $k^{\star}$and the minimum number of $T$. This is because $b$ sets a lower bound on both $k^{\star}$ and $T$. If $k^{\star}$ is small it is less likely that the social norm of kindness breaks down in the first place while a small $T$ reduces the social cost of breakdown. The jointly optimal choice of $T$ and $k^\star$ is beyond the scope of this paper.

## Part 3. Kindness with Persistent Risk

In this chapter we introduce persistence: some agents are more likely than others to be in the position of a helper. First, we consider the case where roles are persistent: with respect to a receiver, an agent who has the role of a helper in one period is more likely to also be a helper in the following period. We call these Markov roles. Notice that, in this case, agents are still homogeneous before the first assignment of roles. In contrast, we then examine persistence in the sense of permanent types. Agents permanently have different likelihoods of being helpers or receivers. Some agents are always more likely to have the role of a helper than others, as in the example of the BMW driver from the introduction. Here, agents are ex-ante heterogeneous. Hence, we first relax the assumption of absence of persistence in agents' roles (Markov roles) and then the assumption of population homogeneity (permanent types).

We show that, for a given expected proportion of receivers in society (i.e., $p$), introducing persistence increases the critical value for $b$ that is necessary to sustain kindness relative to the basic model $\left(b^{Markov}, b^{Types} \geq b^{Base}\right)$. This is because both forms of persistence imply a higher probability of becoming a helper for some agents, relative to the expected probability in society. It is relatively more difficult to convince these agents to be kind, hence they are critical for the computation of the thresholds for $b$. Furthermore, the minimum value for $b$ is higher in the case of permanent types than in the case of Markov roles $\left(b^{Types} \geq b^{Markov}\right)$. In the case of permanent types, some agents always have a higher probability to become a helper, while, in the case of Markov roles, the probabilities reverse with a change of roles.

In the present chapter we will assume a continuum of agents for mathematical convenience, in contrast to chapter 2 where a finite population size was necessary to sustain cooperation in the presence of free-riding.

## 1. Markov Roles

1.1. **Description of the Game.** In this section we relax the assumption that the distribution of future roles is independent of the realization of current roles. The idea is to allow for situations in which being a helper (receiver) in one period implies a different chance of remaining a helper (receiver) in the next period. The structure of the game in the baseline model is maintained with the only difference that now the probability distribution of roles is represented by the following transition matrix:

| | $H_{t+1}$ | $R_{t+1}$ |
|---|---|---|
| $H_t$ | $1 - p_H$ | $p_H$ |
| $R_t$ | $1 - p_R$ | $p_R$ |

where $p_H$, $p_R \in (0, 1)$. We restrict the analysis to the case of a stationary system, where the expected proportion of receivers in the society is constant at $p$. Consequently, given $p_H$, we assume $p_R = (1 - p_H) \cdot \frac{p}{(1-p)}$. The basic model studied previously can be obtained as the special case where $p_H = p_R = p$.

First, we consider a situation where an agent who is a helper in in period $t$ has a lower chance of finding himself in the role of a receiver in period $t+1$ compared to an agent who already is a receiver in period $t$: $p_H \leq p_R$, hence roles are *persistent.* We then compare this case with Hammond's setting of switching roles (Hammond, 1975),[53] where $p_H = 1$, $p_R = 0$ and $p = \frac{1}{2}$. In contrast to the case of persistent roles, in Hammond's model agents switch roles with certainty in every period, hence we have the opposite of persistence.

1.2. **Equilibrium.** In a setting with a continuum of agents who can be identified, a possible strategy to support kindness would be:

> If you are the helper in a relevant match, play K unless the receiver has played NK in a relevant match in the past. In this case play NK.

[53] See literature review for further details on Hammond's model.

In what follows we find conditions for the existence of this cooperative equilibrium. The persistent case does not introduce any strategic novelty with respect to the basic model since a deviator will still receive no further kindness, the only difference being the distribution of types. As a consequence, the existence of a cooperative equilibrium (identical to that proposed in the section about the basic model) is guaranteed by a condition similar to (8), where the differences are uniquely due to the more flexible probability distribution of roles assumed here.[54]

Define $V_{HR}$ as the expected discounted payoff for a helper matched with a receiver if she adheres to the proposed strategy. Define similarly $V_{HH}$, $V_{RR}$ and $V_{RH}$. These can be expressed as:

$$\left.\begin{aligned}
V_{HR} &= -1 + \delta \cdot \{(1-p_H) \cdot [p \cdot V_{HR} + (1-p) \cdot V_{HH}] + p_H \cdot [p \cdot V_{RR} + (1-p) \cdot V_{RH}]\} \\
V_{HH} &= V_{HR} + 1 \\
V_{RR} &= \delta \cdot \{(1-p_R) \cdot [p \cdot V_{HR} + (1-p) \cdot V_{HH}] + p_R \cdot [p \cdot V_{RR} + (1-p) \cdot V_{RH}]\} \\
V_{RH} &= V_{RR} + b
\end{aligned}\right. .$$

As in the basic model, the only agents who have the possibility to deviate are helpers matched with receivers in a given period. Given that everyone else has played K so far, a deviation (NK) is not profitable for a helper in a relevant match, if $V_{HR} \geq 0$. This can be rewritten as:

$$b \geq 1 + \frac{1-\delta}{\delta \cdot p \cdot (1-p)} \cdot \frac{p - \delta \cdot (1-p) \cdot (p - p_H)}{p_H} = b^{Markov}(\delta, p, p_H) \ . \qquad (11)$$

[54] As in the basic model, following the same reasoning, we only consider the collective grim trigger strategy.

$b^{Markov}$ is the minimum value of $b$ necessary to sustain the social norm of kindness with Markov types. It can be easily seen that $b^{Markov} > b^{Base}$ for $p_H < p$. Thus, this form of persistence makes it 'harder' to sustain the social norm of kindness. This makes intuitive sense, as from the perspective of a helper in a given period, it is now less likely to become a receiver in any future period.

1.3. **Hammond's Switching Roles.** In Hammond's model we have the opposite of persistence, since agents switch roles for certain in each period. In our notation $p_H = 1$, $p_R = 0$, and $p = \frac{1}{2}$. After a calculation analogous to the case above, we obtain the critical value (before substituting for $p = \frac{1}{2}$):

$$b^{Switch}\left(\delta, p\right) = 1 + \frac{1-\delta}{\delta \cdot p \cdot (1-p)} \cdot \left[p + \delta \cdot (1-p)^2\right] . \tag{12}$$

It can be easily seen that $b^{Switch}\left(\delta, \frac{1}{2}\right) < b^{Base}\left(\delta, p\right)$, for any $p \neq 0$ and $\delta < 1$.[55] Specifically, for $p = \frac{1}{2}$ we have: $b^{Switch}\left(\delta, \frac{1}{2}\right) = \left(\frac{4}{\delta} - 3\right) \cdot \left(\frac{1}{2} + \frac{\delta}{4}\right)$ and $b^{Base}\left(\delta, \frac{1}{2}\right) = \frac{4}{\delta} - 3$. Thus, the social norm of kindness is 'easier' to sustain with switching roles. This, again, makes intuitive sense since any helper knows that she will be a receiver for sure in the following period, and indeed half of the times she is going to play the game.

## 2. Permanent Types

In this section, we examine a different form of persistence. Here, the likelihood of being helpers or receivers does not depend on the role that was assigned in the last period. Instead, we assume that agents have permanently different likelihoods of being helpers or receivers. Thus, the agents have ex-ante heterogeneous permanent types.

[55] Since $\left[p + \delta \cdot (1-p)^2\right] < 1$.

2.1. **Society-wide Kindness.** An agent's type, denoted by the subscript $i$, is characterized by her likelihood $p_i$ to be a receiver in any given period. In what follows, we refer to this as her personal risk level. The average risk level of an agent in this society is denoted by $p$[56] which is assumed to be common knowledge. Consequently, $p$ can also be interpreted as the expected proportion of receivers in society, and thus as the expected likelihood to be matched with a receiver.

General cooperation (all helpers matched with receivers play K unless they have deviated before) can be maintained if expected payoff from cooperation is positive for all agents. This is the case if the following conditions hold:

$$b \geq \ 1 + \frac{1-\delta \cdot [1-(p-p_i)]}{\delta \cdot p_i \cdot (1-p)} = b_{All}^{Types}\left(\delta, p, p_i\right), \quad \forall i \ .$$

The necessary value for $b$ decreases in the personal risk level $p_i$. Low-risk agents (i.e., agents with a 'small' $p_i$) need a higher benefit to cost ratio in order to be willing to cooperate than high-risk agents (i.e., agents with a 'large' $p_i$). This is so because they are relatively less likely to become receivers, and thus to benefit from the social norm of kindness. Hence, the binding constraint on $b$ is for the agent with the lowest risk level in society:

$$b \geq 1 + \frac{1-\delta \cdot [1-(p-p_{min})]}{\delta \cdot p_{min} \cdot (1-p)} = b_{All}^{Types}\left(\delta, p, p_{min}\right). \tag{13}$$

For any mean preserving spread from a homogeneous society with $p_i = p, \forall i$ to a heterogeneous society with $p_{min} < p$, we have that $b_{All}^{Types} > b^{Base}$. Hence, the required benefit to sustain a general social norm of kindness is higher in a heterogeneous society. This difference is increasing in the degree of heterogeneity as

[56] In order to allow for comparison to the baseline model, it is assumed that the heterogeneous society is obtained through a mean preserving spread from the the baseline model.

given by $(p - p_{min})$. Moreover, the critical value $b_{All}^{Types}$ is unambiguously increasing in $p$. Indeed, with a higher average level of risk in society, a given agent is both relatively less likely to meet a helper when in the role of a receiver, and relatively more likely to meet a receiver when in the role of a helper. Thus, they are more likely to incur a cost than to benefit from the social norm of kindness.

2.2. **Two groups.** Here we consider the case in which agents can have only two types: a share $a \in [0, 1]$ of society has a high-risk of becoming R, $p_h$, while the remaining share $(1 - a)$ has a low-risk of becoming R, $p_l$, where $p_l \leq p_h$. As before we assume this to be a mean preserving spread from the basic case with homogeneous risk $p$, such that: $p = a \cdot p_h + (1 - a) \cdot p_l$. Here, society-wide kindness is beneficial for high-risk agents if:

$$b \geq 1 + \frac{1 - \delta \cdot [1 - (p - p_h)]}{\delta \cdot p_h \cdot (1 - p)} = b_{All}^{h}(\delta, p, p_h) \ , \tag{14}$$

and for low-risk agents if:

$$b \geq 1 + \frac{1 - \delta \cdot [1 - (p - p_l)]}{\delta \cdot p_l \cdot (1 - p)} = b_{All}^{l}(\delta, p, p_l) \ . \tag{15}$$

As stated above, the condition on $b$ which is the hardest to fulfil is for the agent with the lowest risk level in society: $b_{All}^{l} > b_{All}^{h}$ for $p_l < p_h$. So, $b_{All}^{l}$ is the relevant critical value for the feasibility of society-wide cooperation.

*Claim* 6. $b^{Base}(\delta, p) < b_{All}^{l}\left(\delta, p, \underline{p}\right)$, for any $p, \delta, a \in (0, 1)$ and $p_l < p$.[57]

A lower benefit $b$ is always required to make kindness feasible in a homogeneous group with a risk of $p$ than in a heterogeneous society formed through a mean

[57] We obtain $b_{All}^{Types,i} = b^{Base}$, as a special case, when $p_l = p_h = p$.

preserving spread. Hence heterogeneity makes it harder to sustain the social norm of kindness throughout society.

We can also compare this case of permanent types with the case of Markov roles.

*Claim* 7. $b^{Markov}\left(\delta, p, \underline{p}\right) < b^{l}_{All}\left(\delta, p, \underline{p}\right)$, for any $p, \delta, a \in (0,1)$ and $\underline{p} < p$.

A sensible comparison is obtained by setting $p_H = p_l = \underline{p}$. This means that, with Markov roles, the reduced probability of a helper to turn into a receiver in the next period (i.e., $p_H$) is equal, with permanent types, to the reduced probability to be in the position of a receiver for the low-risk types (i.e., $p_l$). Also, to allow for comparison, we consider an average probability to meet a receiver of $p$. In this case, we have:

$$b^{Markov} = 1 + \frac{1-\delta}{\delta \cdot p \cdot (1-p)} \cdot \frac{p - \delta \cdot (1-p) \cdot \left(p - \underline{p}\right)}{\underline{p}} < b^{l}_{All} = 1 + \frac{1 - \delta \cdot \left[1 - \left(p - \underline{p}\right)\right]}{\delta \cdot \underline{p} \cdot (1-p)}.$$

Hence, it is easier to sustain the social norm of kindness with Markov roles than with permanent types and average risk $p$. The intuition for this is that in the case of permanent types, some agents always have a higher probability to become a helper, while, in the case of Markov roles, the probabilities reverse with a change of roles. Thus, from the perspective of an agent with type $p_l$ —the critical agent in the permanent types case—, the probability of being a receiver in any future period is lower than for a helper in the Markov case. But then, the probability of becoming a receiver in the future is the likelihood of —potentially— benefiting from the social norm of kindness. Thus, an agent with a permanently lower risk will be harder to win for cooperation than a helper in the Markov case. Putting this together with the findings from the last section we arrive at the following ranking of critical values for the benefits:

$b^{Switch}\left(\delta,\frac{1}{2}\right) < b^{Base}\left(\delta,p\right) < b^{Markov}\left(\delta,p,\underline{p}\right) < b^{l}_{All}\left(\delta,p,\underline{p}\right)$, for any $p,\delta,a \in (0,1)$ and $\underline{p} < p$.

This ranking is quite intuitive and results from the different degrees of persistence in the different settings. The less persistence, the easier it is to make kindness sustainable. The lowest benefit $b^{Switch}\left(\delta,\frac{1}{2}\right)$ is necessary to sustain kindness in Hammond's switching case, where agents switch every period with certainty. The second lowest benefit, $b^{Base}\left(\delta,p\right)$, is necessary to sustain kindness in a group with a constant, homogeneous risk. The next higher benefit, $b^{Markov}\left(\delta,p,\underline{p}\right)$, is necessary in the Markov case where agents who have the role of receiver in one period are more likely to have this role also in the next period than agents who have the role of helper. Finally, the highest benefit, $b^{l}_{All}\left(\delta,p,\underline{p}\right)$, is necessary in the case with permanent types, where some agents are always more likely than others to be in the role of receivers.

2.3. **Within-Group Kindness.** Even if the social norm of kindness cannot be sustained at the society-wide level, it may still be feasible to sustain it among subgroups of agents with the same personal risk level. The grim-trigger strategy to sustain the social norm of kindness only among agents with the same risk level $p_i$ (i.e., within the high-risk or low-risk group, respectively) is:

> Play K if you are a helper and matched with a receiver **from your group** and if no one **from your group** has defected in the past. When meeting a member of the other group or after a defection in your group, play NK.

For this strategy of exclusive cooperation within groups to work, we assume that agents can observe the risk level of the agent they are matched with and of any potential deviator.[58] If we maintain the assumption that all agents are matched

[58] It could also be an equilibrium for high-risk agents to play K with everyone and for low-risk agents to always play NK. High-risk agents would follow the rule: "Play K if you are a helper and in a relevant match unless another high-risk agent has defected. In that case, play NK forever." Meanwhile, the low-risk agents never cooperate. However, for this strategy to be implemented, high-risk agents would still need to observe the risk-level of a deviator in order to decide whether to stop cooperating or not. Given that types need to be observable in either case,

with equal probability (i.e. irrespective of their types), the probability of being matched with an agent from one's own group is equal to the size of the group as a share of society. If there is cooperation only within one's own group, the members of the group do not receive or exercise kindness when matched with a member of the other group and their payoff is zero in that case. Given this rule, the only agents who have the possibility to deviate are helpers matched with receivers from their group in a given period. In order for the strategy to be an SPNE, their expected payoff from playing K must be positive. Thus, within-group kindness is an equilibrium for the high-risk agents, if:

$$b \geq 1 + \frac{1-\delta}{a \cdot \delta \cdot p_h \cdot (1-p_h)} = b^h_{Group}\left(\delta, a, p_h\right) \ , \tag{16}$$

while for low-risk agents, within-group cooperation is an equilibrium if:

$$b \geq 1 + \frac{1-\delta}{(1-a) \cdot \delta \cdot p_l \cdot (1-p_l)} = b^l_{Group}\left(\delta, a, p_h\right) \ . \tag{17}$$

*Claim* 8. $b^{Base}\left(\delta, p\right) < b^h_{Group}\left(\delta, a, p_h\right), b^l_{Group}\left(\delta, a, p_h\right)$, for any $p, \delta, a \in (0,1)$ and $p_l < p < p_h$.

A lower benefit $b$ is always required to make kindness feasible in a homogeneous group with a risk of $p$ than to support in-group kindness in either the high or low-risk subgroup of a heterogeneous society formed through a mean preserving spread with $a \cdot p_h + (1-a) \cdot p_l = p$. Hence, heterogeneity makes kindness always harder to sustain, be it society-wide as in CLAIM 6, or within groups as here.

Both $b^h_{Group}$ and $b^l_{Group}$ decrease in the share of society of the respective type: $a$ and $(1-a)$. This is so because the greater their share, the more likely it is that

then the proposed strategy of exclusive within-group cooperation, where every agent cooperates but only with those of the same risk level, is welfare dominant.

they are matched with a member of their group and benefit from the social norm of kindness.

*Claim* 9. $b^l_{All} \leq b^l_{Group}$ for any $p_l, p_h, \delta, a \in (0,1)$, $p_l < p_h$ and $\delta \leq \dfrac{1-p_h}{1-p-(1-a)\cdot p_l \cdot (p_h - p_l)}$.

The relationship between $b^l_{All}$ and $b^l_{Group}$ is ambiguous. It might require either a lower or a higher value of $b$ to sustain society-wide kindness for low-risk agents. Specifically, society-wide kindness is easier to sustain than in-group kindness if agents are relatively impatient, i.e., $\delta$ is low enough. This is because a more patient agent is more willing to wait until they meet another agent from their group, while a less patient agent is more willing to include higher risk agents and increase the scope of cooperation. The critical value for $\delta$ decreases in $(p_h - p_l)$, i.e., society wide kindness is easier to achieve if the risk levels are not to different. Moreover, the critical value for $\delta$ is increasing in $a$, i.e., society wide kindness is harder to achieve if the share of high-risk agents is large.

*Claim* 10. $b^l_{Group} \leq b^h_{Group}$ for any $\delta \in (0,1)$ and $(1-a)\cdot p_l \cdot (1-p_l) \geq a \cdot p_h \cdot (1-p_h)$.

Furthermore, also the ranking of $b^l_{Group}$ and $b^h_{Group}$ depends on the parameters. In-group cooperation requires a lower benefit if the type makes up a larger share of the group and for an in-group risk level $p_i$ close to $\frac{1}{2}$. Together, CLAIMS 9 and 10 allow to determine the ordering of $b^l_{All}$,$b^l_{Group}$ and $b^h_{Group}$, in order to determine whether either society-wide kindness or in-group kindness either for low-risk or high-risk agents requires a lower benefit. Depending on the parameters, any ordering is possible.

2.4. **Welfare.** This section explores the welfare ordering of society-wide and in-group cooperation. Specifically, the aggregate expected per-period welfare will be compared for the different settings. For a homogeneous group this would be:

$$W^{Base} = p \cdot (1-p) \cdot (b-1) \, .$$

As before, we compare this to a heterogeneous group with two types that form a mean preserving spread from the homogeneous society. For the case of society-wide kindness, ex-ante expected welfare for the high type is:

$$W^h_{All} = p_h \cdot (1 - p) \cdot b - p \cdot (1 - p_h)$$

and for the low type:

$$W^l_{All} = p_l \cdot (1 - p) \cdot b - p \cdot (1 - p_l) \,.$$

Clearly we have $W^l_{All} < W^h_{All}$ for $p_l < p < p_h$, since in any given period low types are more likely to pay than high types and high types are more likely to receive than low types.

*Claim* 11. $a \cdot W^h_{All} + (1 - a) \cdot W^l_{All} = W^{Base}$ for any $p_l, p_h \in (0, 1)$ and $a \cdot p_h + (1 - a) \cdot p_l = p$.

It can easily be checked that the aggregate welfare from society-wide kindness in a heterogeneous society is exactly the same as the welfare from kindness in the corresponding homogeneous society. Hence, even though a higher benefit is necessary to make a kind equilibrium feasible in a heterogeneous society (see CLAIM 6), if it is feasible the resulting aggregate welfare is the same as in the homogeneous society. However, the welfare gain from kindness is unevenly distributed, since high-risk agents benefit more than low-risk agents.

Next, we look at ex-ante expected welfare for in-group kindness. For the high type this is:

$$W^h_{Group} = a \cdot p_h \cdot (1 - p_h) \cdot (b - 1)$$

and for the low type:

$$W^l_{Group} = (1 - a) \cdot p_l \cdot (1 - p_l) \cdot (b - 1) \, .$$

*Claim* 12. $a \cdot W^h_{Group} + (1 - a) \cdot W^l_{Group} < a \cdot W^h_{All} + (1 - a) \cdot W^l_{All}$ for any $p_l, p_h \in (0, 1)$ and $a \cdot p_h + (1 - a) \cdot p_l = p$.

Comparing the expressions shows that aggregate welfare from in-group kindness is lower than society-wide kindness. Hence, if society-wide kindness is feasible, it is preferred from an aggregate perspective.

*Claim* 13. $W^l_{Group} < W^l_{All}$ for $p_h \cdot (1 - p_l) < b \cdot p_l \cdot (1 - p_h)$.

The low types prefer society-wide to in-group cooperation if the expected gain from extending cooperation to a high type is larger than the expected cost. This is the case when the likelihood of having to pay a high type is smaller than likelihood of receiving from a high type times the value of the benefit $b$. If this is not the case, low agents would prefer in-group kindness, even though society-wide kindness might be feasible and would lead to higher aggregate welfare.

## 3. Conclusion

Chapter 2 shows that the social norm of kindness can arise as an equilibrium of an infinitely repeated dictator game played between randomly matched agents. Neither altruism nor direct reciprocity are necessary to motivate kindness in this setting. Instead, the incentive comes from an uncertainty about one's future position, as agents can find themselves on any side of the dictator game: a receiver in need of kindness or a helper able to show kindness to others. The helper acts kindly, even at a cost, in order to uphold the social norm of kindness which might benefit her in the future when she finds herself in need of kindness. The kindness equilibrium is easiest to sustain when the probability of taking on either role is close to $\frac{1}{2}$.

The social norm is also possible to sustain with imperfect monitoring. In a stylized setting where agents cannot observe every single interaction, but only an aggregate level of kindness, they can use this information to infer the level of compliance in society. There is an equilibrium where agents seize to act kindly if aggregate kindness has fallen short of a specified critical level. The agents act kindly in order to avoid being pivotal for the breakdown of the social norm. This equilibrium becomes harder to sustain as the society grows large. On the other hand, the potential welfare from kindness increases with the size of society.

Chapter 3 shows that the social norm of kindness also becomes harder to sustain when there is persistence in the risk. This is the case for Markov roles, where agents have a higher likelihood of of being in need of kindness in one period if they were in the same position in the previous period. Hence, there is persistence in risk from one period to the next. It becomes even harder if persistence is permanent. In a heterogeneous society where some agents permanently have a higher risk of being receivers than others it might not be possible to sustain a social norm of kindness for the society as a whole. For a sufficiently heterogeneous society, this can lead to a fragmentation of society where kindness is only feasible within more homogeneous subgroups. Kindness is easiest to achieve in large subgroups with a risk level close to $\frac{1}{2}$. Alternatively, kindness might not be sustainable at all in heterogeneous societies.

## Appendix

### Appendix A. Appendix for Chapter 1

A.1. **Derivation of Remark 1.**

$$f\left(\theta_1^i \left| x_1^i \right.\right) = \frac{f_{x_1^i,\theta_1^i}\left(x_1^i,\theta_1^i\right)}{f_{x_1^i}\left(x_1^i\right)} = \frac{f_{x_1^i}\left(x_1^i \left| \theta_1^i \right.\right) f_{\theta_1^i}\left(\theta_1^i\right)}{f_{x_1^i}\left(x_1^i\right)} = \frac{f_\varepsilon\left(x_1^i-\theta_1^i\right) f_{\theta_1^i}\left(\theta_1^i\right)}{f_{x_1^i}\left(x_1^i\right)}$$

$$= \frac{\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{\left(x_1^i-\theta_1^i\right)^2}{2}\right) * \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{\left(\theta_1^i-\mu\right)^2}{2\sigma^2}\right)}{\frac{1}{\sqrt{2\pi(1+\sigma^2)}}\exp\left(-\frac{\left(x_1^i-\mu\right)^2}{2(1+\sigma^2)}\right)} = \frac{1}{\sqrt{2\pi\frac{\sigma^2}{1+\sigma^2}}}\exp\left(-\frac{\left(\theta_1^i-\frac{\mu+\sigma^2 x_1^i}{1+\sigma^2}\right)^2}{2\frac{\sigma^2}{1+\sigma^2}}\right)$$

$$\Rightarrow \left(\theta_1^i \left| x_1^i \right.\right) \sim N\left(\frac{\mu+\sigma^2 x_1^i}{1+\sigma^2}, \frac{\sigma^2}{1+\sigma^2}\right)$$

A.2. **Proof of proposition 1.** In the following the superscript $i$ will be suppressed. In order to show that equation 5

$$F_{\theta_K}\left(z\left|x_1 \le \mu\right.\right) - F_{\theta_S}\left(z\left|x_1 \le \mu\right.\right) \ge 0 \quad \forall z,$$

holds, both terms are rewritten:

$$F_{\theta_S}\left(z\left|x_1 > \mu\right.\right) = F_{\theta_S^i}(z) = \int_{-\infty}^{z} f_{\theta_1}(t)\, dt$$

.

$$\begin{aligned}
f_{\theta_K}\left(z\left|x_1 \le \mu\right.\right) &= \frac{f_{\theta_K}(z)\, \cdot\, \Pr\left(x_1 \le \mu\left|z\right.\right)}{\Pr\left(x_1 \le \mu\right)} \\
&= \frac{f_{\theta_K}(z)\, \cdot\, F_\varepsilon(\mu - z)}{F_{x_1}(\mu)} \\
&= 2\, f_{\theta_1}(z)\ F_\varepsilon(\mu - z), \\
\Rightarrow F_{\theta_K}\left(z\left|x_1^i \le \mu\right.\right) &= \int_{-\infty}^{z} 2\, f_{\theta_1}(t)\ F_\varepsilon(\mu - t)\, dt.
\end{aligned}$$

Substituting for $F_{\theta_S}\left(z\left|x_1 \le \mu\right.\right)$ and $F_{\theta_K}\left(z\left|x_1 \le \mu\right.\right)$, 5 becomes:

$$\int_{-\infty}^{z} f_{\theta_1}(t)\left[2\, F_\varepsilon(\mu - t) - 1\right]\, dt > 0\, \forall z$$

For all $s$ the two following properties hold: $f_{\theta_1}(\mu - s) = f_{\theta_1}(\mu - s)$ , since $f_{\theta_1}$ is symmetric around $\mu$, and $F_\varepsilon(-s) = 1 - F_\varepsilon(-s)$, since $f_\varepsilon$ is symmetric around 0.

For any $z \leq \mu$ we have that $\int_{-\infty}^{z} F_\varepsilon(\mu - t)\, dt > \frac{1}{2}$ so that $\int_{-\infty}^{z} f_{\theta_1}(t)\left[2\, F_\varepsilon(\mu - t) - 1\right] dt > 0\ \forall z \leq \mu$.

For any $z > \mu$,

$$\int_{-\infty}^{z} f_{\theta_1}(t)\left[2\, F_\varepsilon(\mu - t) - 1\right] dt$$

equals

$$\begin{aligned} &\int_{-\infty}^{2\mu - z} f_{\theta_1}(t)\left[2\, F_\varepsilon(\mu - t) - 1\right] dt \\ &+ \int_{2\mu - z}^{\mu} f_{\theta_1}(t)\left[2\, F_\varepsilon(\mu - t) - 1\right] dt \\ &+ \int_{\mu}^{z} f_{\theta_1}(t)\left[2\, F_\varepsilon(\mu - t) - 1\right] dt \end{aligned}$$

The second integral on this expression, with the change of variable $s = \mu - t$, becomes:

$$\int_{z-\mu}^{0} f_{\theta_1}(\mu - s)\left[2\, F_\varepsilon(s) - 1\right] ds = \int_{0}^{z-\mu} f_{\theta_1}(\mu + s)\left[2\, F_\varepsilon(s) - 1\right] ds$$

where we use the symmetry of $f_{\theta_1}$. The third integral, using $s = t - \mu$ and the symmetry of $F_\varepsilon$, is:

$$\begin{aligned}
&\int_{0}^{z-\mu} f_{\theta_1}(\mu+s)\,[2\,F_\varepsilon(-s)-1]\;ds \\
=&\int_{0}^{z-\mu} f_{\theta_1}(\mu+s)\,\{2\;[1-F_\varepsilon(s)]-1\}\;ds \\
=&\int_{0}^{z-\mu} f_{\theta_1}(\mu+s)\,[1-2\,F_\varepsilon(s)]\;ds
\end{aligned}$$

Thus the second and third term sum to zero. It follows that:

$$\int_{-\infty}^{z} f_{\theta_1}(t)\,[2\,F_\varepsilon(\mu-t)-1]\;dt\;=\int_{-\infty}^{2\mu-z} f_{\theta_1}(t)\,[2\,F_\varepsilon(\mu-t)-1]\;dt.$$

For any $z>\mu$ we have that $\int_{-\infty}^{2\mu-z} F_\varepsilon(\mu-t)\;dt>\frac{1}{2}$ so that $\int_{-\infty}^{z} f_{\theta_1}(t)\,[2\,F_\varepsilon(\mu-t)-1]\;dt>0\;\forall z>\mu$.

Hence, 5 and thus proposition 1 holds. The agent will follow the recommendation 'S' of the principal when $x_1^i\leq\mu.\square$

A.3. **Proof of proposition 2.** In the following the superscript $i$ will be suppressed. In order to show that equation 6

$$F_{\theta_S}(z\,|x_1>\mu)-F_{\theta_K}(z\,|x_1>\mu)\geq 0\;\forall z$$

holds, both terms are rewritten:

$$F_{\theta_S}(z\,|x_1>\mu)=F_{\theta_S}(z)=\int_{-\infty}^{z} f_{\theta_1^a}(t)\;dt$$

$$\begin{aligned}
f_{\theta_K}(z\,|x_1>\mu)\;&=\;\frac{f_{\theta_K}(z)\;\cdot\;\Pr(x_1>\mu\,|z)}{\Pr(x_1>\mu)}\\
&=\;\frac{f_{\theta_K}(z)\;\cdot\;[1-F_\varepsilon(\mu-z)]}{1-F_{x_1}(\mu)}\\
&=\;2\,f_{\theta_1^a}(z)\;F_\varepsilon(z-\mu)\\
\Rightarrow F_{\theta_K}\left(z\left|x_1^i\leq\mu\right.\right)\;&=\;\int_{-\infty}^{z} 2\,f_{\theta_1}(t)\;F_\varepsilon(t-\mu)\;dt.
\end{aligned}$$

Substituting for $F_{\theta_S}(z\,|x_1 > \mu)$ and $F_{\theta_K}(z\,|x_1 > \mu)$, 6 becomes:

$$\int_{-\infty}^{z} f_{\theta_1^a}(t)\left[1 - 2\,F_{\varepsilon}(t-\mu)\right]\,dt > 0\;\forall z$$

Since $1 - 2\,F_{\varepsilon}(t-\mu) = 2\,F_{\varepsilon}(\mu - t) - 1$ the proof in A.2 applies from here. Hence, 6 and thus proposition 2 holds. The agent will follow the recommendation 'K' of the

## Appendix B. Appendix for Chapter 2

B.1. **Proof of CLAIM 2.** By De Moivre-Laplace Theorem, as $N$ grows large while $q$ is constant, $\Pr\left(K_{N-1} = k^{\star} - 1|\,q\right)$ can be approximated in a neighbourhood of $q \cdot (N-1)$ by $\frac{1}{\sqrt{2\cdot\pi\cdot(N-1)\cdot q\cdot(1-q)}} \cdot \exp\left\{\frac{-[k^{\star}-1-q\cdot(N-1)]^2}{2\cdot(N-1)\cdot q\cdot(1-q)}\right\}$, a Normal density function. The maximum of this Normal density function (and of the probability mass function that it is approximating) is achieved in $k^{\star}-1 = q\cdot(N-1)$ and equals $\frac{1}{\sqrt{2\cdot\pi\cdot(N-1)\cdot q\cdot(1-q)}}$. As $N \to +\infty$, the maximum of the Normal density function approaches 0. Also, for any value of the parameters, $1 > \delta \cdot \Pr\left(K_N \geq k^{\star}|\,q\right) \geq 0$. The result follows. ■

B.2. **Proof of CLAIM 3.** Compute the derivative:

$$\frac{\partial b^{Free}(\delta, N, q, k^\star)}{\partial q} < 0$$

$$\Leftrightarrow -\delta \cdot \frac{\partial \Pr(K_N \geq k^\star)}{\partial q} \cdot \left[\frac{\delta \cdot k^\star}{2 \cdot N} \cdot \Pr(K_N = k^\star)\right]^{-1}$$

$$- \left[1 - \delta \cdot \Pr(K_N \geq k^\star)\right] \cdot \left[\frac{\delta \cdot k^\star}{2 \cdot N} \cdot \Pr(K_N = k^\star)\right]^{-2}$$

$$\cdot \frac{\delta \cdot k^\star}{2 \cdot N} \cdot \frac{\partial \Pr(K_N = k^\star)}{\partial q} < 0$$

$$\Leftrightarrow \overbrace{-\delta \cdot \left\{\sum_{x=k^\star}^{N} \Pr(K_N = x) \cdot [(1-q) \cdot x + N \cdot q - x \cdot q]\right\}}^{LHS} < \overbrace{[1 - \delta \cdot \Pr(K_N \geq k^\star)] \cdot [(1-q) \cdot k^\star + N \cdot q}^{RHS}$$

where the second equivalence follows from:

$$\frac{\partial \Pr(K_N = k^\star)}{\partial q} = \left[\frac{k^\star - 2 \cdot q \cdot k^\star + N \cdot q}{q \cdot (1-q)}\right] \cdot \Pr(K_N = k^\star)$$

$$.$$

$$\frac{\partial \Pr(K_N \geq k^\star)}{\partial q} = \sum_{x=k^\star}^{N} \Pr(K_N = x) \cdot \left[\frac{(1-q) \cdot x + N \cdot q - x \cdot q}{q \cdot (1-q)}\right]$$

We need to show that $\frac{\partial b^{Free}(\delta,N,q,k^\star)}{\partial q} < 0$ or, equivalently, that $LHS < RHS$. Notice how $(1-q) \cdot x + N \cdot q - x \cdot q > 0$, $\forall q > 0$. Indeed, $(1-q) \cdot x - x \cdot q \in [0, x)$ for $q \in (0, 0.5]$. Hence, $RHS$ is always positive and $LHS$ always negative. ■

## References

Abramitzky, R., 2008. The limits of equality: Insights from the israeli kibbutz. The quarterly journal of economics 123 (3), 1111–1159.

Alesina, A., La Ferrara, E., 2000. Participation in heterogeneous communities. The Quarterly Journal of Economics 115 (3), 847–904.

Alesina, A., La Ferrara, E., 2002. Who trusts others? Journal of Public Economics 85 (2), 207–234.

Ambrus, A., Mobius, M., Szeidl, A., 2010. Consumption risk-sharing in social networks.

Andreoni, J., 1989. Giving with impure altruism: Applications to charity and ricardian equivalence. The Journal of Political Economy, 1447–1458.

Andreoni, J., 1990. Impure altruism and donations to public goods: a theory of warm-glow giving. The Economic Journal 100 (401), 464–477.

Andreoni, J., 1995. Cooperation in public-goods experiments: kindness or confusion? The American Economic Review, 891–904.

Angelucci, M., De Giorgi, G., 2009. Indirect effects of an aid program: How do cash transfers affect ineligibles' consumption? The American Economic Review 99 (1), 486–508.

Aoyagi, M., 2010. Information feedback in a dynamic tournament. Games and Economic Behavior 70 (2), 242–260.

Arad, A., Rubinstein, A., 2013. Strategic tournaments. American Economic Journal: Microeconomics 5 (4), 31–54.

Arcand, J.-L., Fafchamps, M., 2012. Matching in community-based organizations. Journal of Development Economics 98 (2), 203–219.

Athey, S., Calvano, E., Jha, S., 2010. A theory of community formation and social hierarchy. Unpublished manuscript.

Attanasio, O., Barr, A., Cardenas, J. C., Genicot, G., Meghir, C., 2012. Risk pooling, risk preferences, and social networ. American Economic Journal: Applied Economics 4 (2), 134–167.

Banerjee, S., Konishi, H., Sönmez, T., 2001. Core in a simple coalition formation game. Social Choice and Welfare 18 (1), 135–153.

Bardsley, N., 2008. Dictator game giving: altruism or artefact? Experimental Economics 11 (2), 122–133.

Barr, A., Genicot, G., 2008. Risk sharing, commitment, and information: an experimental analysis. Journal of the European Economic Association 6 (6), 1151–1185.

Bloch, F., Genicot, G., Ray, D., 2007. Reciprocity in groups and the limits to social capital. The American Economic Review 97 (2), 65–69.

Bloch, F., Genicot, G., Ray, D., 2008. Informal insurance in social networks. Journal of Economic Theory 143 (1), 36–58.

Bogomolnaia, A., Jackson, M., 2002. The stability of hedonic coalition structures. Games and Economic Behavior 38 (2), 201–230.

Bold, T., 2009. Implications of endogenous group formation for efficient risk-sharing. The Economic Journal 119 (536), 562–591.

Bolton, G. E., Ockenfels, A., 2000. Erc: A theory of equity, reciprocity, and competition. The American economic review, 166–193.

Camerer, C., 2003. Behavioral game theory: Experiments in strategic interaction. Princeton University Press.

Charness, G., Genicot, G., 2009. Informal risk sharing in an infinite-horizon experiment*. The Economic Journal 119 (537), 796–825.

Charness, G., Rabin, M., 2002. Understanding social preferences with simple tests. The Quarterly Journal of Economics 117 (3), 817–869.

Cherry, T. L., Frykblom, P., Shogren, J. F., 2002. Hardnose the dictator. The American Economic Review 92 (4), 1218–1221.

Choy, J., 2013. A theory of cooperation through social division, with evidence from nepal.

Coate, S., Ravallion, M., 1993. Reciprocity without commitment: Characterization and performance of informal insurance arrangements. Journal of development Economics 40 (1), 1–24.

Dana, J., Weber, R. A., Kuang, J. X., 2007. Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness. Economic Theory 33 (1), 67–80.

DellaVigna, S., List, J. A., Malmendier, U., 2012. Testing for altruism and social pressure in charitable giving. Tech. Rep. 1.

Dreze, J., Greenberg, J., 1980. Hedonic coalitions: Optimality and stability. Econometrica: Journal of the Econometric Society, 987–1003.

Dufwenberg, M., Kirchsteiger, G., 2004. A theory of sequential reciprocity. Games and Economic Behavior 47 (2), 268–298.

Ederer, F., 2010. Feedback and motivation in dynamic tournaments. Journal of Economics & Management Strategy 19 (3), 733–769.

Ederer, F., August 2013. Incentives for parallel innovation. Working Paper.

Ellison, G., 1994. Cooperation in the prisoner's dilemma with anonymous random matching. The Review of Economic Studies 61 (3), 567–588.

Fafchamps, M., 2008. Risk sharing between households. Handbook of Social Economics, 1–42.

Fafchamps, M., Gubert, F., 2007. The formation of risk sharing networks. Journal of Development Economics 83 (2), 326–350.

Fafchamps, M., Lund, S., 2003. Risk-sharing networks in rural philippines. Journal of development Economics 71 (2), 261–287.

Fehr, E., Schmidt, K. M., 1999. A theory of fairness, competition, and cooperation. The quarterly journal of economics 114 (3), 817–868.

Fowler, J. H., Christakis, N. A., 2010. Cooperative behavior cascades in human social networks. Proceedings of the National Academy of Sciences 107 (12), 5334–5338.

Fuchs, W., 2006. Contracting with repeated moral hazard and private evaluations. American Economic Review 97 (4).

Genicot, G., Ray, D., 2003. Group formation in risk-sharing arrangements. The Review of Economic Studies 70 (1), 87–113.

Gershkov, A., Perry, M., 2009. Tournaments with midterm reviews. Games and Economic Behavior 66 (1), 162–190.

Gertler, P., Gruber, J., 2002. Insuring consumption against illness. The American Economic Review 92 (1), 50–70.

Ghosh, P., Ray, D., 1996. Cooperation in community interaction without information flows. The Review of Economic Studies 63 (3), 491–519.

Glazer, A., Hassin, R., 1988. Optimal contests. Economic Inquiry 26 (1), 133–143.

Goltsman, M., Mukherjee, A., 2011. Interim performance feedback in multistage tournaments: The optimality of partial disclosure. Journal of Labor Economics 29 (2), 229–265.

Gomes, R., Gottlieb, D., Maestri, L., December 2013. Experimentation and project selection: Screening and learning. Working Paper.

Gouldner, A. W., 1960. The norm of reciprocity: A preliminary statement. American sociological review, 161–178.

Green, E. J., Porter, R. H., 1984. Noncooperative collusion under imperfect price information. Econometrica: Journal of the Econometric Society, 87–100.

Green, J. R., Stokey, N. L., 1983. A comparison of tournaments and contracts. Journal of Political Economy 91 (3), 349–64.

Griffith, P., Norman, W., O'Sullivan, C., Ali, R., 2011. Charm offensive: Cultivating civility in 21st century britain. The Young Fundation.

Grossman, P. J., Eckel, C. C., 2012. Giving versus taking: A ”real donation” comparison of warm glow and cold prickle in a context-rich environment.

Halac, M. C., Liu, Q., Kartik, N., November 2012. Optimal contracts for experimentation. Working Paper.

Hammond, P., 1975. Charity: Altruism or cooperative egoism. Altruism, Morality and Economic Theory. New York: Russell Sage Foundation.

Hoffman, E., McCabe, K., Smith, V. L., 1996. Social distance and other-regarding behavior in dictator games. The American Economic Review 86 (3), 653–660.

Jalan, J., Ravallion, M., 1999. Are the poor less well insured? evidence on vulnerability to income risk in rural china. Journal of development economics 58 (1), 61–81.

Kandori, M., 1992. Social norms and community enforcement. The Review of Economic Studies 59 (1), 63–80.

Karlan, D., Mobius, M., Rosenblat, T., Szeidl, A., 2009. Trust and social collateral. The Quarterly Journal of Economics 124 (3), 1307–1361.

Kimball, M. S., 1988. Farmers' cooperatives as behavior toward risk. The American Economic Review 78 (1), 224–232.

Kini, O., Williams, R., 2012. Tournament incentives, firm risk, and corporate policies. Journal of Financial Economics 103 (2), 350–376.

Kletzer, K. M., Wright, B. D., 2000. Sovereign debt as intertemporal barter. The American Economic Review 90 (3), 621–639.

Kocherlakota, N. R., 1996. Implications of efficient risk sharing without commitment. The Review of Economic Studies 63 (4), 595–609.

Konrad, K. A., 2009. Strategy and Dynamics in Contests. Oxford University Press.

Kremer, I., Mansour, Y., Perry, M., May 2013. Implemeting the wisdom of the crowd. Working Paper.

Lazear, E. P., Rosen, S., 1981. Rank-order tournaments as optimum labor contracts. The Journal of Political Economy, 841–864.

Leahy, T., 2012. Management in 10 Words. Cornerstone Digital.

Leider, S., Möbius, M. M., Rosenblat, T., Do, Q.-A., 2009. Directed altruism and enforced reciprocity in social networks. The Quarterly Journal of Economics 124 (4), 1815–1851.

Levine, D. K., 1998. Modeling altruism and spitefulness in experiments. Review of economic dynamics 1 (3), 593–622.

Ligon, E., Schechter, L., 2011. Motives for sharing in social networks. Journal of Development Economics.

Ligon, E., Thomas, J. P., Worrall, T., 2000. Mutual insurance, individual savings, and limited commitment. Review of Economic Dynamics 3 (2), 216–246.

List, J. A., 2007. On the interpretation of giving in dictator games. Journal of Political Economy 115 (3), 482–493.

Lizzeri, A., Meyer, M. A., Persico, N., August 2002. The incentive effects of interim performance evaluations. Unpublished Manuscript.

Manso, G., 2011. Motivating innovation. The Journal of Finance 66 (5), 1823–1860.

Mauss, M., 1924. Essai sur le don forme et raison de l'échange dans les sociétés archaïques. L'Année sociologique 1, 30–186.

Meyer, M. A., 1991. Learning from coarse information: Biased contests and career profiles. The Review of Economic Studies 58 (1), 15–41.

Miguel, E., Gugerty, M. K., 2005. Ethnic diversity, social sanctions, and public goods in kenya. Journal of Public Economics 89 (11), 2325–2368.

Moscarini, G., Squintani, F., 2010. Competitive experimentation with private information: The survivor's curse. Journal of Economic Theory 145 (2), 639–660.

Murgai, R., Winters, P., Sadoulet, E., Janvry, A. d., 2002. Localized and incomplete mutual insurance. Journal of Development Economics 67 (2), 245–274.

Nalebuff, B. J., Stiglitz, J. E., 1983. Prizes and incentives: towards a general theory of compensation and competition. Bell Journal of Economics 14 (1), 21–43.

Nowak, M. A., 2012. Evolving cooperation. Journal of Theoretical Biology 299, 1–8.

Nowak, M. A., Sigmund, K., 1998. Evolution of indirect reciprocity by image scoring. Nature 393 (6685), 573–577.

Nowak, M. A., Sigmund, K., 2005. Evolution of indirect reciprocity. Nature 437 (7063), 1291–1298.

Okuno-Fujiwara, M., Postlewaite, A., 1995. Social norms and random matching games. Games and Economic Behavior 9 (1), 79–109.

Rabin, M., 1993. Incorporating fairness into game theory and economics. The American Economic Review, 1281–1302.

Rohner, D., 2011. Reputation, group structure and social tensions. Journal of Development Economics 96 (2), 188–199.

Rosenthal, R. W., 1979. Sequences of games with varying opponents. Econometrica: Journal of the Econometric Society, 1353–1366.

Rosenthal, R. W., Landau, H. J., 1979. A game-theoretic analysis of bargaining with reputations. Journal of Mathematical Psychology 20 (3), 233–255.

Seinen, I., Schram, A., 2006. Social status and group norms: Indirect reciprocity in a repeated helping experiment. European Economic Review 50 (3), 581–602.

Selten, R., Ockenfels, A., 1998. An experimental solidarity game. Journal of economic behavior & organization 34 (4), 517–539.

Sigmund, K., 2012. Moral assessment in indirect reciprocity. Journal of Theoretical Biology 299, 25–30.

Townsend, R. M., 1994. Risk and insurance in village india. Econometrica: Journal of the Econometric Society, 539–591.

Udry, C., 1994. Risk and insurance in a rural credit market: An empirical investigation in northern nigeria. The Review of Economic Studies 61 (3), 495–526.

Watson, J., 1999. Starting small and renegotiation. Journal of economic Theory 85 (1), 52–90.

Weynants, S., 2011. Informal insurance with endogenous group size.

Ythier, J. M., Kolm, S.-C., Gerard-Varet, L.-A., 2001. The economics of reciprocity, giving and altruism. Vol. 130. Palgrave Macmillan.