CEDEX

CENTRE FOR DECISION RESEARCH & EXPERIMENTAL ECONOMICS

The University of Nottingham

UNITED KINGDOM · CHINA · MALAYSIA

Discussion Paper No. 2014-02

Simon Gächter

March 2014

Human Pro-Social Motivation and the Maintenance of Social Order

CeDEx Discussion Paper Series

ISSN 1749 - 3293

CENTRE FOR DECISION RESEARCH & EXPERIMENTAL ECONOMICS

The Centre for Decision Research and Experimental Economics was founded in 2000, and is based in the School of Economics at the University of Nottingham.

The focus for the Centre is research into individual and strategic decision-making using a combination of theoretical and experimental methods. On the theory side, members of the Centre investigate individual choice under uncertainty, cooperative and non-cooperative game theory, as well as theories of psychology, bounded rationality and evolutionary game theory. Members of the Centre have applied experimental methods in the fields of public economics, individual choice under risk and uncertainty, strategic interaction, and the performance of auctions, markets and other economic institutions. Much of the Centre's research involves collaborative projects with researchers from other departments in the UK and overseas.

Please visit http://www.nottingham.ac.uk/cedex for more information about the Centre or contact

Suzanne Robey
Centre for Decision Research and Experimental Economics
School of Economics
University of Nottingham
University Park
Nottingham
NG7 2RD
Tel: +44 (0)115 95 14763
Fax: +44 (0) 115 95 14159
suzanne.robey@nottingham.ac.uk

The full list of CeDEx Discussion Papers is available at

http://www.nottingham.ac.uk/cedex/publications/discussion-papers/index.aspx

HUMAN PRO-SOCIAL MOTIVATION AND THE MAINTENANCE OF SOCIAL ORDER

Simon Gächter*, University of Nottingham, CESifo, IZA

16 March 2014

Forthcoming in: Eyal Zamir and Doron Teichman, *Handbook on Behavioral Economics and the Law*, Oxford University Press

Abstract

This chapter presents some insights from basic behavioural research on the role of human pro-social motivation to maintain social order. I argue that social order can be conceptualised as a public good game. Past attempts to explain social order typically relied on the assumption of selfish and rational agents ("homo economicus"). The last twenty years of research in behavioural and experimental economics have challenged this view. After presenting the most important findings of recent research on human pro-sociality I discuss the evidence on three pillars of the maintenance of social order. The first pillar is internalised norms of cooperation, sustained by emotions such as guilt and shame. The second pillar is the behaviour of other people who typically are "conditional cooperators" willing to cooperate if others do so as well. This motivation can sustain cooperation if enough people cooperate but can jeopardise social order if many others follow selfish inclinations. The third pillar are sanctions meted out to anyone who does not cooperate; ideally punishment can work as a mere threat without being executed much. The chapter also presents some evidence on the cross-cultural variability of some findings, in particular with regard to punishment behaviour. The chapter concludes with remarks on future research.

Keywords: social order; social dilemma; pro-sociality; strong reciprocity; moral emotions; social norms; conditional cooperation; punishment; rule of law.

* University of Nottingham, School of Economics, Sir Clive Granger Building, University Park, Nottingham NG7 2RD, United Kingdom. Email: simon.gaechter@nottingham.ac.uk. Support under the European Research Council grant ERC-AdG 295707 COOPERATION is gratefully acknowledged. I thank Doron Teichman and Eyal Zamir for their comments.

1. INTRODUCTION

Economic approaches to understanding human behavior, including law abidance, have long assumed that people are self-regarding in the sense that they entertain cost-benefit calculations with the sole concern being own costs and benefits, irrespective of consequences for others. The last twenty years of research in behavioral economics have profoundly challenged this assumption (Gintis, Bowles, Boyd, and Fehr 2005) with important consequences for our understanding of lawful behavior and social order in general. The question I will discuss in this chapter is how pro-social motivations help understand social order.

I will discuss evidence from the last two decades of behavioral economics research that sheds light on human pro-social motivations. I will focus my attention on people's behavior in social settings where the welfare of other people is affected. Of course, there are important behavioral aspects of law abidance from an asocial, individual decision-making perspective. These concern the roles of probability perception (e.g., the perceived probability of being caught for a criminal act) and heuristics and biases in general (Kahneman and Tversky 2000). I do not deal with these issues here but refer the interested reader to Sunstein (2000), and to the chapter on heuristics and biases by Baron in this volume. My focus is on pro-social motivation, not cognition.

The basic conceptual framework I will use to study social order is going back to at least Hobbes and consists in thinking of social order as a cooperation problem: If the law is widely disregarded we end up in a world were life is "nasty, brutish, and short". Contributing to social order (obeying the law) is of collective interest but individuals have an incentive to disregard the law if this promises to be more advantageous than abiding by the law. Of course, the law also has an important coordinating function which makes obeying the law also in people's self-interest - think of traffic laws, for instance. In this chapter, however, I will not discuss coordination problems, but focus on cooperation.

One may argue that modern societies rest on constitutions and legal enforcement mechanisms with checks and balances that sharply limit individual decisions to flout the law. However, as examples from failed states vividly demonstrate, legal enforcement cannot work if large groups in society disregard the law because they think respecting the law is not to their advantage. Thus, a functioning law enforcement system is itself an example of successful cooperation.

The problem of cooperation can most easily be understood in the following simple example. Suppose two farmers reach an agreement to respect each other's property but there is no third party to enforce this agreement. Now the two farmers have to think whether to stick to their promise or not. Suppose they both abide by their agreement and therefore have an incentive to cultivate their land which gives them a comfortable living. But one farmer might be tempted to renege on his promise and steal the harvest of the other farmer who trusted that he would be safe and therefore invested in a good crop. The stealing farmer enjoys a harvest for which he did not work, and the victim is robbed of his proceeds. If farmers are not gullible (or learn from experience) they might anticipate this outcome and not invest much in cultivating the land which leaves both in a miserable situation but still better than losing all harvest after a season of hard work. This, of course, is the famous prisoners' dilemma: mutually sticking to the agreement is in the common interest but not in each individual's interest.

The prisoners' dilemma as a metaphor for cooperation has been the focus of decades of research (Rapoport and Chammah 1965; Axelrod 1984; Van Lange, Balliet, Parks, and Van Vugt 2014). One important insight has been that cooperation (i.e., honoring the agreement) might be maintained in my example if these farmers are likely to play the game in the foreseeable future with the "shadow of the future" (Axelrod 1984) looming strongly enough. The mutual threat to renege on the agreement if the other farmer reneges might be strong enough to honor the agreement. If successful, we have an example of self-enforcement. Such self-enforcement is much harder to achieve if the players are not settled farmers but mobile bands of hunter-gatherers because under the latter conditions there is no common shadow of the future but only a short-term cooperation problem, which favors defection.

Modern social life differs of course from this simple example: cooperation problems need to be solved for large groups, where decisions take place both in stable settings and random interactions. But large groups, even if they are stable, are fundamentally different from two-person prisoner's dilemmas: theory suggests (Boyd and Richerson 1988) and experimental evidence confirms (Grujić et al. 2012) that stable cooperation is possible in the two-person prisoner's dilemma but is hard to achieve in large groups because no effective punishment targeted at non-compliant group members exists. Thus, for understanding large-scale cooperation the prisoner's dilemma is not fully suitable and recent research has therefore shifted to the public goods game as a tool to study multi-lateral cooperation. This game will be the major tool I will use in this chapter.

One important insight of many experiments using the public goods game is that cooperation is inherently unstable and tends to unravel to the worst outcome, predicted by self-interest. Doesn't this prove that people are selfish in the end? My answer will be a qualified No. Some people are indeed likely to be selfish. Many people, however, will *behave* selfishly under some conditions, but are not *motivated* by selfishness. As I will show, the distinction between motivation and behavior is important and ought not to be conflated. People can be non-selfishly motivated and end up behaving selfishly, but the converse also exists: selfish people behaving pro-socially.

The main tool to investigate my questions has been economic experiments, with decision-dependent monetary stakes. A full description of the experimental methodology is beyond the scope of this chapter. The interested reader should consult Falk and Heckman (2009) and Engel (in this volume).

The plan of this chapter is as follows. Section 2 will lay the foundation of my analysis of determinants of social order by offering an overview of the most important findings suggesting that the *homo economicus* assumption used for decades in economics and other behavioral sciences is not justified. Many people are more aptly described as *homo reciprocans*, i.e., non-selfish "strong reciprocators" (Gintis 2000) and I will present the most important evidence supporting the existence of strong reciprocators. A strong reciprocator is prepared to sacrifice resources to be kind to those who are being kind ("strong positive reciprocity") and to punish those who are being unkind ("strong negative reciprocity"). The essential feature of strong reciprocity is a willingness to reward fair and punish unfair behavior even if this is costly and provides neither present nor future material rewards for the reciprocator (Fehr, Fischbacher, and Gächter 2002). However, as I will show, all experiments that find evidence for strong reciprocity also find the existence of mostly self-regarding people.

The rest of this chapter will then discuss how *homo economicus* and *homo reciprocans* deal with social order. I will argue that social order is sustained, to some extent, by internalized norms of proper conduct even in the absence of any formal enforcement. Social order is also, and very strongly so, influenced by the behavior of other people because *homo reciprocans* is more likely to contribute to the common good if others do the same. I will also show that punishment or other incentives are necessary to sustain social order.

A first pillar of social order, and probably the weakest one, is personal ethics, or internalized norms of cooperation, enforced by feelings of guilt: Cooperation can be supported to the extent that people think cooperating is the morally right thing to do and feel guilty if breaking the social contract. Section 3 investigates the role of internalized norms of proper conduct to sustain cooperation. In Section 4 I will show that social order is bound to be fragile if not backed up by incentives. This holds despite the fact that most people are not fundamentally self-regarding and, as Section 3 will show, express moral apprehension at free riding. An important insight is that some people are selfish and that *homo reciprocans*, while not being selfish, sometimes tends to be selfishly biased. Section 5 discusses evidence that (the threat of) punishment is crucial to maintain social order. *Homo reciprocans* has a decisive role to play because *homo reciprocans* is prepared to pay a cost to punish those who jeopardize social order. Rewards and a desire for a good reputation can also help.

Section 6 will present some cross-societal evidence and show that punishment is also shaped by how well the Rule of Law works in a given society. Section 7 will present a short discussion and outlook for future research.

Before I proceed I should clarify what this chapter does and does not provide. Research in the behavioral sciences searches for basic behavioral principles that underlie all social dilemmas however diverse they are in reality. My approach therefore is not applied science (although I will point to some interesting applied findings) but basic science that should provide general behavioral principles that can inform more applied research. The behavioral research I report here is complementary to approaches studying the role of social norms in law and its enforcement (e.g., Ellickson 1991; Posner 2000; Kahan 2003).

2. BASIC SOCIAL MOTIVATIONS: *HOMO ECONOMICUS* AND *HOMO RECIPROCANS*

Homo economics has long been the most important characterization of human nature in the behavioral sciences and in particular in economics. David Hume famously remarked that "Political writers have established it as a maxim, that, in contriving any system of government, and fixing the several checks and controls of the constitution, every man ought to be supposed a knave, and to have no other end, in all his actions, than private interest" (Hume 1987 [1777], Essay VI, p. 42). George Stigler, a Nobel laureate in Economic

Sciences, was convinced: “Let me predict the outcome of the systematic and comprehensive testing of behavior in situations where self-interest and ethical values with wide verbal allegiance are in conflict. Much of the time, most of the time in fact, the self-interest theory … will win" (Stigler 1981, p. 175).

There are several justifications for the selfishness assumption. *Homo economicus* is neutral to other people, that is, he is neither envious or malicious and also not altruistic. Thus, he might be considered the average person on whom social analysis should be based (Kirchgässner 2008). Furthermore, in a theoretical context, the *homo economicus* assumption often allows for exact predictions, which can be confronted with appropriate data that might refute it. Moreover, it is often of independent interest to understand what would happen if everyone were self-regarding. A clear picture of the consequences of selfishness serves therefore as an important benchmark for understanding non-selfish behavior.

The assumption of self-regard also has considerable merit in the absence of empirical means to assess the structure of people’s social preferences. Yet, the experimental methodology allows us to observe people’ social preferences under controlled circumstances. Advances in neuroscience (Glimcher, Camerer, Fehr, and Poldrack 2009), anthropology (Henrich et al. 2004), behavioral economics (Gintis et al. 2005), evolutionary theory (Bowles and Gintis 2011) and social psychology (Van Lange et al. 2014) shed further light on human nature. Thus, given the availability of appropriate tools to measure deviations from selfishness, there is no need to rely further on the selfishness assumption. Its empirical relevance can now be measured.

In the following I will present evidence that supports the widespread existence of *homo reciprocans*. The classic games used to study people’s social preferences are the dictator game, the ultimatum game, the trust and the gift-exchange game, and the public goods game with and without punishment. All experiments I will discuss are conducted according to the standards of experimental economics (see Friedman and Sunder 1994 for a textbook account) and have been replicated many times, including in representative samples, under high stakes, and in relevant field conditions. Moreover, all experiments are designed to carefully control for self-regarding incentives, such that self-interest theory makes a unique prediction that can be compared with the behavioral outcome. If behavior differs from the self-interest prediction we have evidence for non-selfish behavior.

The *dictator game* (Forsythe, Horowitz, Savin, and Sefton 1994) is the most basic decision situation in which social preferences can be studied. The dictator game is a two-player game where participants are assigned at random to be either a "dictator" or a passive recipient. The dictator has to decide how much of a given amount of money allocated to him or her to share with a recipient who has to accept the offer. The experimental setting ensures that a self-interested rational dictator has an incentive not to share. Passing money along to the recipient under these conditions is evidence for altruism, or other-regarding preferences in general.

The results of many carefully controlled dictator games do not support self-interest predictions on average. In a meta-analysis Engel (2011) finds that across 616 treatments involving the dictator game, the average sharing rate is 28.3 percent and across all studies about 36 percent of individuals do not share at all. Thus, many people are willing to share a windfall gain, but (depending on the treatment) a sizeable minority is not.

How about sharing principles if recipients can reject the offer? The seminal game to study this situation is the *ultimatum* game (Güth, Schmittberger, and Schwarze 1982). In the ultimatum game the proposer makes an offer of how to share a given pie and, in contrast to the dictator game, the recipient can now accept or reject the offer. In case of acceptance, the offered division is implemented; in case the recipient rejects, both get nothing. If the recipient is motivated solely by monetary payoffs, he or she will accept every offer. Therefore, the proposer will only offer the smallest money unit.

The results across a wide range of subject pools around the world reject this prediction (Oosterbeek, Sloof, and van de Kuilen 2004). On average, proposers offer 30 to 40 percent of the available amount. The median and the mode are at 40 and 50 percent, respectively. Few offers are less than10 percent, or more than 50 percent. Offers below 20 percent or less will likely be rejected, while equal splits are almost always accepted.

The offers made in the ultimatum game appear inconsistent with the *homo economicus* model of human nature. However, it is important to observe that all types of proposers, self- and other-regarding ones, have an incentive to offer non-minimal (fair) shares, if some recipients are inclined to reject low offers. Thus the mere fact that we observe high offers is not inconsistent with the *homo economicus* model. The inconsistency arises for the recipient who foregoes earnings by rejecting a positive offer – *homo economicus* would never do that. Cross-societal variation notwithstanding, when it comes to rejections, there is abundant

support for the existence of strong negative reciprocity, and no support for the *homo economicus* prediction in almost any of the many societies studied (Oosterbeek et al. 2004; Henrich et al. 2005; Henrich et al. 2006).

The next game, the *gift-exchange game* (developed by Fehr, Kirchsteiger, and Riedl 1993), showcases strong positive reciprocity where *homo reciprocans* non-strategically rewards a kind act by being kind as well. A simple version of the gift-exchange game works as follows. There are two roles, employers and employees. In each round, an employer and employee are paired up at random. The employer makes a wage offer to his or her paired employee, who can accept or reject the offer. Acceptance concludes an employment contract. The employee then chooses effort and the round ends. "Effort" means choosing a number with the consequence that the higher the chosen number the higher is the employer's profit and the higher are the employee's effort cost. The earnings of employers increase in effort and decrease in wages paid. For the employee the opposite holds. Parameters are such that maximal effort maximizes surplus.

The setup ensures that there are no strategic reasons for gift exchange. A *homo economicus* employee will choose the minimum effort irrespective of the wage because effort is costly. *Homo reciprocans*, however, will respond reciprocally: high wages are rewarded with high effort and low wages are matched with low effort.

The results of numerous experiments support the *homo reciprocans* prediction over the *homo economicus* one, on average, because wage and effort are highly significantly correlated. This is unambiguous evidence for strong positive reciprocity, found in numerous gift-exchange experiments (see Charness and Kuhn 2011 for an overview). However, the results also reveal substantial heterogeneity. Irrespective of the wage paid by the firm there is always a fraction of workers who choose minimal effort – like in the dictator game *homo economicus* exists but is in the minority.

A game related to the gift-exchange game that also allows for the observation of strong positive reciprocity is the *trust game* (Berg, Dickhaut, and McCabe 1995). The trust game is a two-player game where participants are anonymously and at random allocated to their roles as trustor and trustee. The trustor (and in some experiments also the trustee) has an endowment and has to decide how much of this endowment to transfer to the trustee. Any amount the trustor transfers the experimenter increases by a factor of 3 (in some studies by a factor of 2 or 4). The trustee then decides how much of the increased amount to transfer back

to the trustor. *Homo economicus* in the role of recipient will not return anything irrespective of the amount received (and rational trustors would foresee this and transfer nothing).

Numerous studies with student and wide-ranging non-student subject pools have been conducted with the trust game. Johnson and Mislin (2011) found in a meta-analysis of 162 replications in 35 countries that trustors on average send 50 percent of their endowment and trustees return 37 percent of the amount available for return. Regression analyses show clear support for strong positive reciprocity: Trustees return highly significantly more the more they have received from the trustor.

In sum, the gift-exchange game and the trust game provide substantial evidence for the existence of strong positive reciprocity. Rejections in the ultimatum game are an example of strong negative reciprocity. But these are all two-player games. I have argued in the introduction that to understand human cooperation one needs to move beyond dyadic interactions. The following game, the public good game, is a vehicle to study strong positive reciprocity in the context of a simultaneous multi-lateral game.

In a typical *linear public good game*, four people form a group. All group members are endowed with 20 tokens. Each member has to decide independently how many tokens (between 0 and 20) to contribute to a common project (the public good). The contributions of the whole group are summed up. The experimenter then multiplies the sum of contributions by a factor larger than one but less than four (a frequently used factor is 1.6) and distributes the resulting amount equally among the four group members irrespective of how much an individual has contributed. Thus, an individual benefits from the contributions of other group members, even if he or she has contributed nothing to the public good. A rational and self-regarding individual has an incentive to keep all tokens, because the personal benefit per token from the public good is less than one, whereas it is 1 if he or she keeps the token. By contrast, the group as a whole is best off if everybody contributes all 20 tokens.

A large number of studies show that people contribute to the public good (see Chaudhuri 2011 for an overview), but, as I will describe in more detail in the next section, contributions decrease over time in experiments that allow for repetition of the base game. In this section I focus on one-shot games because my goal is to demonstrate the existence of strong positive reciprocity, and this requires controlling for any self-regarding incentives. One-shot games provide the starkest environment to study cooperation motivated by strong reciprocity because there are no strategic reasons to make any positive contribution. Thus, *homo*

economicus will take a free ride in this game. Again, the results do not confirm this prediction. Many people make a positive contribution, although a significant fraction contributes nothing (e.g., Dufwenberg, Gächter, and Hennig-Schmidt 2011).

The fact that people make positive contributions does not yet constitute evidence for strong positive reciprocity. In a game where group members make their contribution decisions simultaneously people cannot respond to what they have observed others to do; people can only react to the *beliefs* they hold about other group members' contributions. Thus, some experiments ask the participants what they estimate the other group members will contribute (e.g., Dufwenberg et al. 2011). The results are consistent with strong positive reciprocity: on average, reported beliefs and own contributions are highly significantly positively correlated. While this holds for a majority of people, some contribute nothing despite the fact they believe others will contribute a lot. Again, *homo economicus* and *homo reciprocans* co-exist.

The final game I discuss is the *public good game with punishment* (Fehr and Gächter 2000; Fehr and Gächter 2002). It provides another example for the existence of strong negative reciprocity. In this game, the group members first contribute to the public good. Then group members learn how much all have contributed and are given the opportunity to spend money to reduce the income of each of the other group members individually. One money unit spent on punishing a group member reduces this group member's earnings from the first stage by three money units. *Homo economicus* will of course not spend any money to punish others, but *homo reciprocans* might be willing to punish the free riders in the group.

The results show that many people are prepared to punish free riders. In fact, in the experiments of Fehr and Gächter (2002) more than 80 percent of people punished at least once. Fehr and Gächter repeated their experiment six times, but each time with entirely new group members and in a way that excluded any further interactions with any previous group members. Punishment showed a reciprocal pattern in each of the six one-shot repetitions: more free riding was met with more punishment. Gächter and Herrmann (2009) and Cubitt, Drouvelis, and Gächter (2011a) found the same result in strict one-shot experiments. Such punishment has been called "altruistic" because it is individually costly and benefits others only; it is evidence of strong negative reciprocity. These experiments and a related large literature (surveyed in Gächter and Herrmann 2009; Balliet, Mulder, and Van Lange 2011;

and Chaudhuri 2011) show that many people are strong negative reciprocators with punitive sentiments for wrongdoing.

What are possible psychological (proximate) mechanisms that produce strong reciprocity? At the most fundamental level, it is the evolved human capacity of empathy which only psychopaths lack (Baron-Cohen 2011). Relevant for my specific question, research has identified three important mechanisms: inequality aversion (Fehr and Schmidt 1999; Bolton and Ockenfels 2000); efficiency seeking (Charness and Rabin 2002); and a desire to reward or punish intentions behind actions (also called reciprocity; Rabin 1993; Dufwenberg and Kirchsteiger 2004) or a combination of inequality aversion and rewarding and punishing intentions (Falk and Fischbacher 2006). A detailed description is beyond the scope of this chapter, but I will provide the ideas. Thorough discussions can be found in the cited articles and in Fehr and Schmidt (2006); see Wilkinson and Klaes (2012) for a textbook account. Bowles and Gintis (2011) provide evolutionary (ultimate) explanations of strong reciprocity.

Inequality aversion. Inequality aversion, in particular the version by Fehr and Schmidt (1999), is probably the most widely used theory to explain the reviewed behavior in experimental games. The theory assumes that people care about their own material payoff positively and negatively about inequality in comparison with another person both in case the inequality is advantageous (the focal individual has more than the comparison individual) and if it is disadvantageous (the focal individual has less than the comparison individual). It is assumed that disadvantageous inequality is worse than advantageous inequality.

The theory of inequality aversion can explain why people reject unfair offers in ultimatum games (an example of strong negative reciprocity): while there is a positive utility from the material benefit of a (small) offer, there is also disutility from inequality. It is therefore possible that the disutility outweighs the utility from the offered amount, and therefore total utility is negative and the person rejects. A second example of how disadvantageous inequality aversion can explain strong negative reciprocity is punishment of free riders in a public good game: a free rider who does not contribute anything will earn more than all others who have contributed. This will leave the contributors behind in payoff comparisons and they will experience disadvantageous inequality aversion. A contributor who punishes a free rider may reduce the gap in earnings and therefore inequality by punishing. Aversion to advantageous inequality can also explain why people behave in a

positive reciprocal way when making contributions to a public good. If a group member believes others will contribute he or she might feel advantageous inequality aversion if not contributing. To alleviate this feeling, she contributes. An inequality averse person will also not contribute more than others because this way she would fall behind in terms of payoffs.

There are, however, a couple of important phenomena that the theory of inequality aversion does not address: many people are motivated by efficiency seeking and are therefore willing to help others even if that increases inequality (a strictly inequality averse person would not do that), and people care not only about outcomes as assumed in theories of inequality aversion, but also about the intentions behind actions. I will deal with these two problems in turn.

Efficiency seeking. Inequality aversion implies that people will always take actions, if available, that reduce inequality. But experiments have shown that many people are also willing to help other people if that increases efficiency despite also increasing inequality (Charness and Rabin 2002; Engelmann and Strobel 2004). Thus, a social *concern for efficiency* most likely is an important motivation for some, and it might also explain why people make contributions to public goods.

Intentions matter. A second problem with the theory of inequality aversion is that it is purely outcome-oriented, i.e., the *intentions* behind other people's actions do not matter. However, there are many cases where intentionality is important. For example, receiving an unfair offer (involving a disadvantageous unequal distribution) if a fair offer (an equal distribution) is available might not be perceived the same as if the only available other offer is also unfair (Falk, Fehr, and Fischbacher 2003). Theories of reciprocity (e.g., Dufwenberg and Kirchsteiger 2004) model intentions by assuming that people are motivated by rewarding kindness with kindness and meanness by meanness. Making a fair offer when an unfair offer is available (and better for the proposer) is an example of a kind act; offering an unfair distribution when a fair one would have been available is an example of unkind behavior. Another example is contributions to a public good: if a group member believes others will contribute a lot then he or she might perceive this as a kind act and reward the kindness by contributing as well; by the same token, a low expected contribution might be perceived as unkind and therefore be matched with a low contribution as well (Dufwenberg et al. 2011). Falk and Fischbacher (2006) combine inequality aversion and intentions, and show that

intentions might lead to more punishment of unfair offers and free riding than inequality aversion alone.

In sum, strong positive and negative reciprocity are probably to a large extent motivated by psychological mechanisms of inequality aversion and a desire to base rewards and punishments on the intentions behind an action; in some important cases concerns for social efficiency also matter. Existing research clearly suggests an important role for these mechanisms (Falk, Fehr, and Fischbacher 2005). For my purposes, however, it suffices to work with strong positive and negative reciprocity as motivational shortcuts.

In the following I will turn to the central question of this chapter, the determinants of social order. In the next three sections I will show how the basic inclinations of strong positive and negative reciprocity determine (the breakdown of) social order.

3. THE DETERMINANTS OF SOCIAL ORDER I: INTERNALIZED NORMS

One important determinant of people's pro-social behavior is most likely internalized norms of what people consider the morally right thing to do. For example, people donate anonymously to charities (Eckel and Grossman 1996), they vote for reasons of civic duty, despite their vote being extremely unlikely pivotal (Riker and Ordeshook 1968); they respect the law (Cooter 2000) also if incentives that back up the obligations are weak (Galbiati and Vertova 2008). People pay their taxes despite low detection probabilities for evasion (Kirchler 2007), and people also care for the environment out of moral convictions (Brekke, Kipperberg, and Nyborg 2010). More generally, people value character virtues such as honesty and trustworthiness even if lying and cheating go entirely undetected (e.g., Gneezy 2005; López-Pérez and Spiegelman 2013; see Pruckner and Sausgruber 2013 for an interesting field study) and also act on perceived moral obligations (Schwartz 1977). As shown above, in experiments people make contributions to one-shot public goods without any extrinsic incentive to do so. One early piece of evidence that is consistent with intrinsic motivations is that people contribute for reasons of "warm glow" (Andreoni 1990).

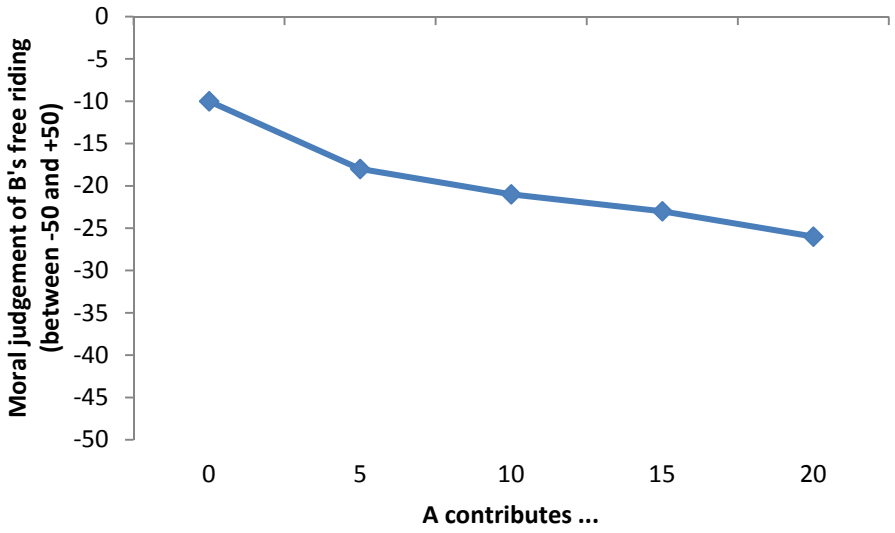
In this section I discuss some evidence about normative considerations and related moral emotions in social dilemmas. I discuss studies that investigate people's moral judgments, the

social emotions of anger and guilt, and people's desire to punish norm violators even if not personally affected ("third-party punishment").

I start with moral judgments of free riding. Is free riding morally blameworthy at all? Cubitt, Drouvelis, Gächter, and Kabalin (2011b) report on a study that elicited people's moral judgments of free riding by using techniques from moral psychology to understand to what extent free riding is perceived to be a *moral* problem. The basic design of Cubitt and his colleagues' study is as follows. They presented their subjects – who took the roles of spectators – with scenarios of two people, A and B, who are both endowed with 20 money units and make contributions to a public good. B always free rides, that is, keeps all of his 20 money units for himself. The different scenarios vary the extent to which A makes contributions to the public good. Depending on the scenario, A contributes 0, 5, 10, 15 or 20 to the public good. People were asked, as a detached observer, how they morally judge B's behavior for each of A's possible contribution. The moral judgment scale ranged from -50 (extremely bad) to +50 (extremely good).

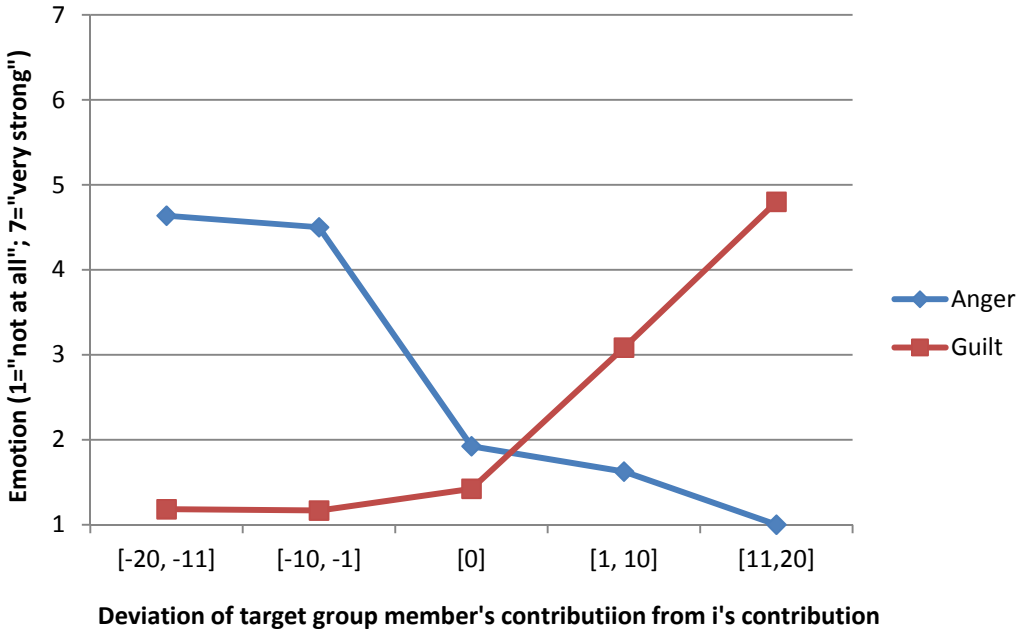
Figure 1A illustrates the result by showing the average moral evaluation of B's free riding (contribution of 0 to the public good) for each of A's possible contribution. The average moral evaluation is always below 0, that is, people think that B's free riding is morally blameworthy. Interestingly, the same act of free riding is considered morally worse on average the more A actually contributes.

Figure 1A: Moral judgment of free riding



Data source: Cubitt et al. (2011b); own illustration.

Figure 1B: Moral emotions – anger and guilt



Data source: Cubitt et al. (2011a); own illustration.

Figure 1A shows the average moral evaluation, which hides some interesting heterogeneity. About 50 percent of people actually have a flat "moral judgment function" that is, their moral evaluation of B's free riding does not depend on how much A contributes. A third of the people think B's free riding becomes morally worse the more A contributes.

If free riding is considered morally blameworthy, does it also trigger negative emotions? Evidence using non-involved spectators who evaluate free riding behavior described in various scenarios, suggest so (Fehr and Gächter 2002). And if cooperation is morally commendable, does free riding trigger feelings of guilt? Anger and guilt are expected to be particularly relevant in a context of social cooperation because they can be seen as prototypical morally-linked emotions (e.g., Haidt 2003).

Cubitt et al. (2011a) elicited emotions after players made their contributions in a one-shot public good to see to what extent free riding triggers anger by the cooperating individual and guilt by the free rider. Figure 1B shows that the average levels of anger and guilt seem to be mirror images of one another with the exception that a high level of free riding (where the target individual contributes between 11 and 20 tokens less than the focus individual) triggers the same anger as a lower level of free riding (the target individual contributes between 1 and 10 tokens less than the focus individual).

The moral or social emotions anger and guilt are interesting because they trigger two potential enforcement mechanisms – external and internal punishment. Angry individuals might be willing to punish free riders and therefore provide the free riders with an extrinsic self-regarding incentive to avoid punishment by contributing (discussed in more detail in Section 6). Guilt is a negative emotion that can serve as "internal punishment" and therefore provide an intrinsic reason to contribute to the public good to avoid feeling guilty. Dufwenberg et al. (2011) presented evidence that such "guilt aversion" can explain contributions to public goods.

Further evidence for the importance of normative considerations comes from third-party punishment games (Fehr and Fischbacher 2004), where a potential punisher is not an affected party, but an independent third party (this feature thus resembles law enforcement in reality). In their experiment, two players, A and B, play a Prisoner's Dilemma game with two options: Cooperate (C) or Defect (D). In terms of material payoffs, the best outcome for a player is DC, that is to defect when the other player cooperates; the second-best outcome is CC (mutual cooperation); the third-best result is DD (mutual defection) and the worst outcome is

CD (cooperating while the other player defects). This incentive structure gives both players an incentive to defect and therefore to forego the gains from mutual cooperation. Fehr and Fischbacher (2004) add to this framework a third party who, at own cost, can punish both players A and B after having seen their decisions. Since the third party is not affected by A and B's decisions, third-party punishment is a reflection of normative considerations. The results show that third parties are much more likely to punish a defector if the other player cooperated (in 46 percent of cases) than if both defected (21 percent of cases); mutual cooperation is almost never punished.

The results on third-party punishment are consistent with the findings on moral judgments (Figure 1A): free riding is considered particularly blameworthy if the other party cooperated. The third-party experiments uncovered that people have a willingness to pay for their normative convictions. Neuro-scientific evidence (Buckholtz and Marois 2012) as well as cross-cultural findings (Henrich et al. 2006) suggest that third-party punishment is a phenomenon that is deeply ingrained in the human condition.

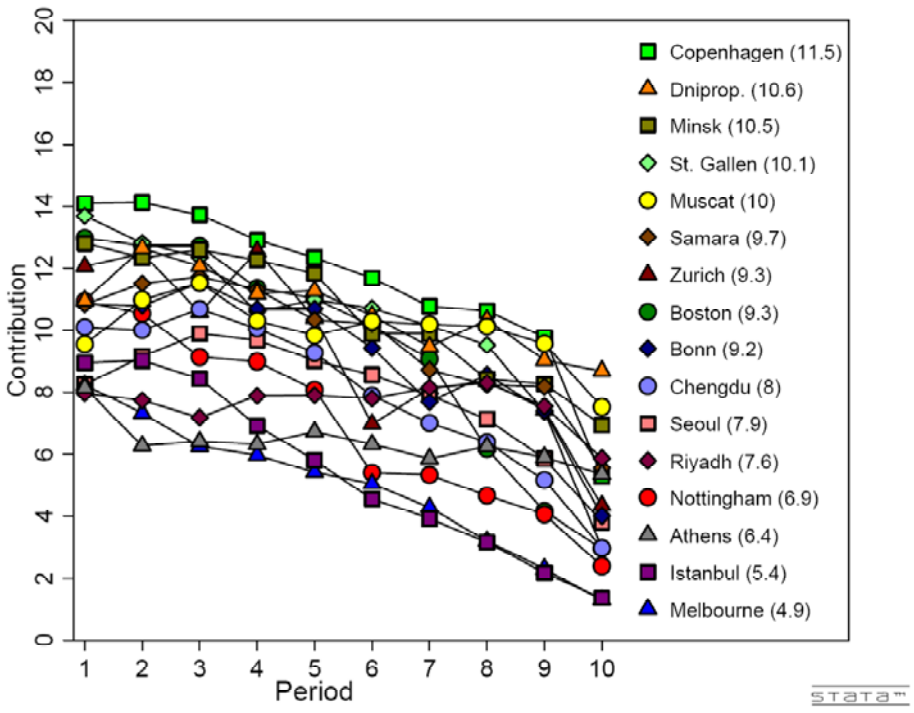
In summary, people think free riding is morally blameworthy and it also triggers the contributors' anger and even third-party punishment. People who contribute less than others feel guilty. Thus, to the extent that people have feelings of warm glow, are bound by moral norms, want to avoid making other group members angry even if (third-party) punishment is not possible, and would feel guilty if contributing less than others, pro-social cooperation is expected.

4. THE DETERMINANTS OF SOCIAL ORDER II: THE BEHAVIOR OF OTHER PEOPLE

I introduced the *one-shot* public good game in Section 2 as one tool to study the existence of strong positive reciprocity. The evidence suggests that people are willing to contribute to public goods even in one-shot settings. To investigate (the stability of) social order, however, requires repeated public good games. Notice that the repeated public good game is a stark setting to study social order: while one-shot settings allow observing people's principle willingness to cooperate for the sake of the collective benefit, a repeated setup allows answering the question whether this willingness can help producing a *stable* social order.

Are people able to provide a public good which has a collective benefit to all, if the collective benefit and the "shadow of the future" are the sole incentives? The fact that many people are guilt-averse, think free riding is immoral, and are also motivated by efficiency-seeking should help in pursuing collective welfare. However, the sobering answer of many repeatedly played public good experiments is that cooperation almost invariably breaks down in repeated interactions. This result has been shown in numerous experiments around the world (Herrmann, Thöni, and Gächter 2008; Chaudhuri 2011) and is illustrated in Figure 3. In all subject pools people contribute substantial amounts initially but over time contributions dwindle to low levels almost everywhere.

Figure 2: The breakdown of cooperation is ubiquitous: Evidence from fifteen countries



Source: Herrmann et al. (2008) (Figure 3). Figure 2 shows the average contribution (out of the endowment of 20 money units) the subjects contributed in each round. The numbers in parentheses are average contributions over all rounds.

One may argue that the experiments reported in Figure 2 are too short to properly reflect conditions relevant for social order. Unfortunately, existing experimental evidence suggests that the time horizon does not matter much. For example, in Gächter, Renner, and Sefton

(2008) participants played for ten or fifty periods (and participants new this). Cooperation was low under both time horizons (less than 40 percent on average) but not different between time horizons. Rand et al. (2009) and Grujić et al. (2012) report very similar results (Grujić et al. even for 100 periods, and like in Rand et al. 2009 with participants being unaware of the exact number of rounds). Thus, the conclusion is inevitable and seems to vindicate Hobbes: in and of itself, that is, without external enforcement, social order is fragile and the time horizon as such is of no avail.

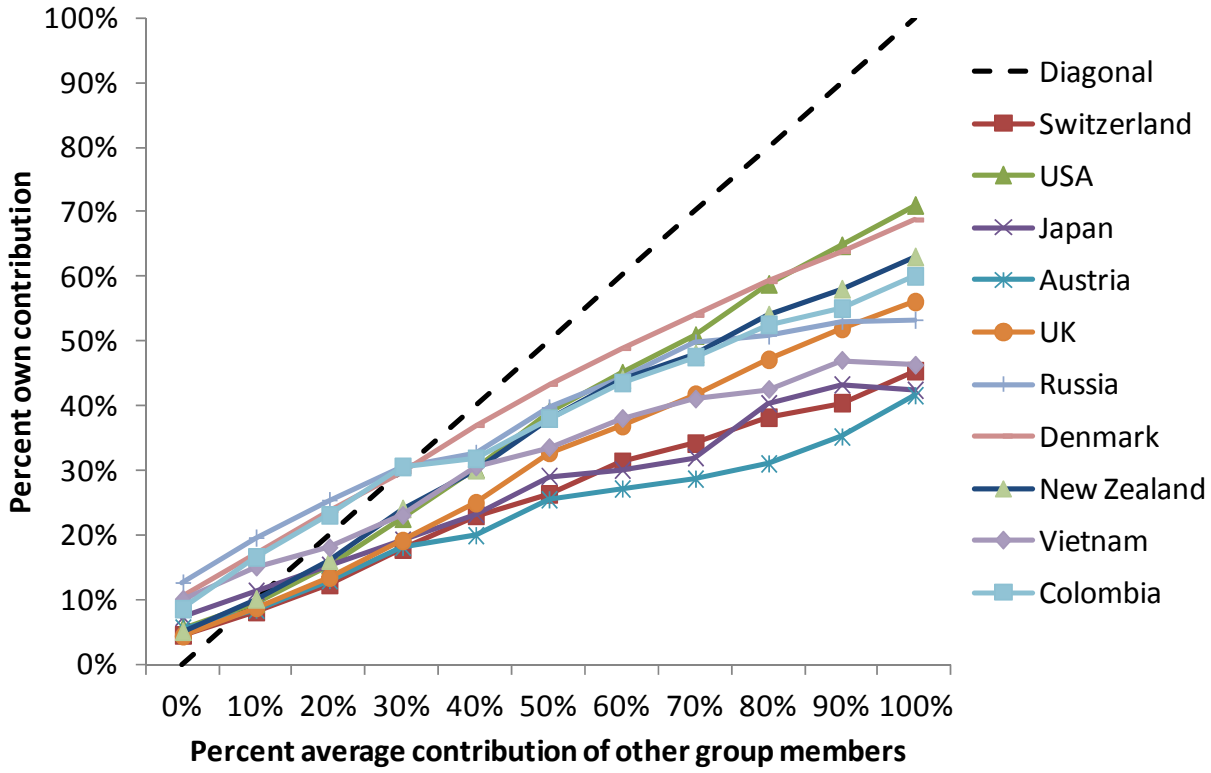
Recall from Section 2 that one-shot public good experiments have found a positive correlation between beliefs about the contributions of other group members and own contributions, which is consistent with strong positive reciprocity. However, this positive correlation is not a particularly compelling measure of strong positive reciprocity. To see why, suppose, for whatever reason, Alice is very pessimistic about the contributions of others and thinks they will not contribute much or even nothing at all. Suppose further Alice would be willing to contribute provided others also contribute – Alice is a "conditional cooperator". Alice behaves as a free rider due to her pessimism not because her basic attitude to cooperation is free riding. Now compare Alice to Bill and assume that Bill is a free rider who will never contribute even if others contribute a lot. Thus, there is a problem: Alice and Bill both free ride, so their *behavior* is observationally equivalent, but their *motivation* is different. Bill is motivated to be a free rider, whereas Alice is a conditional cooperator who happens to be pessimistic. Thus, separating behavior from motivation is important (see Lewinsohn-Zamir 1998 for a related argument in a law and public policy context).

Fischbacher, Gächter, and Fehr (2001) introduce a design that allows separating behavior from motivation. In their experiment, participants are asked in an incentive-compatible way to make conditional contributions for all possible average contributions of the other group members (a so-called "strategy method"). Given the details of the incentive structure, people motivated by free riding will contribute nothing for all levels of possible average contributions of other group members. Conditional cooperators, by contrast, will increase their contribution in the average contribution of others. Thus, in this design, rather than just observing one contribution and one belief, we can observe a complete contribution schedule for all possible average contributions of others. Free riders and conditional cooperators are therefore clearly distinguishable, even if they both contribute nothing if others contribute nothing. Fischbacher et al. (2001) find that about 50 percent of their participants are

conditional cooperators, 30 percent are free riders, and the rest follow some other patterns. The average person clearly is a conditional cooperator.

The Fischbacher et al. (2001) experiment has been replicated many times in many countries and including representative subject pools (Thöni, Tyran, and Wengström 2012). Figure 3 illustrates the average conditional contribution from subjects in ten different countries around the world by showing the average contribution that subjects make as a function of all possible average contribution levels of other group members (expressed in percentages of the maximal possible contribution which differs across studies).

Figure 3: The average person is a conditional cooperator: Evidence from ten countries



Data source: Chaudhuri and Paichayontvijit (2006) (New Zealand); Kocher et al. (2008) (Austria, Japan, USA); Herrmann and Thöni (2009) (Russia); Fischbacher et al. (2012) (UK, Switzerland); Thöni et al. (2012) (Denmark); Martinsson, Pham-Khanh, and Villegas-Palacio (2013) (Vietnam, Colombia). Own illustration.

A couple of important insights can be taken away from Figure 3. First, although there is some variation, patterns are very similar across subject pools: low contributions by other

group members are met with low own contributions, and own contributions increase in those of group members. This is true in all ten subject pools illustrated here. Second, while contributions increase in the contributions of others, own contributions tend to remain below the diagonal, which implies that even conditional cooperators on average want to free ride to some extent on the contributions of other group members. Figure 3 depicts average conditional cooperation and it therefore hides heterogeneity. However, conditional cooperators are the majority and free riders a minority in all subject pools studied. The assumption that *homo economicus* describes average behavior is thus not supported by the experimental findings.

Before I move on to discuss how the observation of Figure 3 can explain the fragility of social order, it is worth discussing three more observations about conditional cooperation: several psychological mechanisms predict conditional cooperation which makes it a highly likely pattern; conditional cooperation is externally valid; and conditional cooperation predicts contributions in experimental public good games.

Several psychological mechanisms support conditional cooperation. Conditional cooperation is a likely pattern of behavior because various psychological mechanisms predict it. I already mentioned two proximate mechanisms of strong reciprocity in Section 2 – *inequality aversion* and a desire to match like with like (*reciprocity*). Numerous experiments suggest the existence of inequality aversion and reciprocity and I have already sketched the argument how these motivations can explain conditional cooperation. Conditional cooperation is also supported by cooperative *social value orientations* where people take into account the welfare of others (Balliet, Parks, and Joireman 2009; Van Lange et al. 2014). A further channel to support conditional cooperation is *guilt aversion*, introduced in Section 3. If Alice thinks others expect her to contribute she might feel guilty if she wouldn't and to avoid feeling guilty she actually makes a contribution to the public good; if she expects others not to contribute, she will also not feel guilty by not contributing herself.

Moreover, conformism, a deep-rooted human tendency to copy other people's behavior, also supports conditional cooperation. A desire to conform will lead a conformist to contribute if he or she thinks that is what other people will do; of course conformists will also free ride if that is what the majority does. This argument has found some experimental support (Carpenter 2004).

Conditional cooperation has external validity. Conditional cooperation is not only observed under laboratory conditions but also in naturally occurring environments. For example, field experiments demonstrate donations to public goods consistent with conditional cooperation (e.g., Frey and Meier 2004; Shang and Croson 2009). Rustagi, Engel, and Kosfeld (2010) ran experiments with forest management groups in Ethiopia. They employ a measure similar to that used in the experiments summarized in Figure 3 and show that groups with a high share of conditional cooperators are more successful in forest management (an important public good in Ethiopia) than groups with a higher share of free riders. A final example is tax morale, which displays the behavioral logic of conditional cooperation that is, people are more likely to be honest in their tax declaration if they think most other people are as well (Frey and Torgler 2007; Traxler 2010).

Conditional cooperation predicts contributions. Conditional cooperation is not only a phenomenon with high external validity; it is also internally valid in the sense that the elicited cooperation preferences predict actual play in new public goods games: people classified as conditional cooperators also behave as conditional cooperators in a new public good game and free riders tend to contribute nothing as predicted for them (Fischbacher, Gächter, and Quercia 2012). Moreover, when attitudes to cooperation are elicited multiple times, most people fall into the same type categorization each time, that is, conditional cooperation and free riding are intra-personally stable attitudes (Volk, Thöni, and Ruigrok 2012). This observation supports evidence that people's other- or self-regarding behavior is consistent across games (Yamagishi et al. 2013).

These observations are important for explaining why social order in and of itself, that is, without further incentives, is inherently fragile. As Figure 3 shows, the average person is a conditional cooperator, but detailed analyses show that some people are free rider types who never contribute. Moreover, on average, even conditional cooperators are selfishly biased. Most conditional cooperators will make a positive initial contribution to the public good and then take the average contribution of the other group members as the new benchmark. The fact that most conditional contributors are also selfishly biased will induce them to contribute less than the average next time and therefore cooperation will almost inevitably unravel and finally most people will contribute little or nothing to the public good. This prediction is consistent with the evidence (see Figure 2; Fischbacher and Gächter 2010 for a rigorous analysis; and Chaudhuri 2011 for a survey of this literature).

This result of the unraveling of cooperation due to selfishly-biased conditional cooperation teaches us two important lessons. First, due to the process of conditionally cooperative reactions on others' contribution, many people will eventually behave like a free rider (contribute little to the public good) despite the fact that they are not motivated by selfishness. Second, cooperation is inherently fragile, and needs some support through other mechanisms to be sustainable.

One assumption I have been making so far is that people are sorted at random into groups, and all experiments I discussed did in fact implement random group assignment. However, in reality, people can sometimes choose the social group they want to be in. Thus, the question is, does sorting help? The answer is a qualified yes. If people manage to sort into groups with strongly reciprocal conditional cooperators then such groups are indeed able to maintain high levels of cooperation and can prevent its breakdown (Gächter and Thöni 2005). This observation is consistent with conditional cooperation: if others cooperate conditional cooperators will cooperate too. But successful sorting requires that the cooperative types are indeed sorted together and are able to prevent free riders from entering (Ehrhart and Keser 1999) and can credibly signal their type (for a discussion of signaling from a law point of view see Posner 2000). These are quite stringent conditions that may or may not be satisfied in real social groups.

In summary, conditional cooperation is an important human motivation for many and, as numerous experiments have shown, a highly relevant determinant of social order. Thus, although conditional cooperation allows for the *possibility* of self-sustaining cooperation, it is unlikely that conditional cooperators manage to maintain high levels of cooperation. This is due to the existence of a substantial fraction of free riders and to the fact that even conditional cooperators typically display some selfish bias.

5. THE DETERMINANTS OF SOCIAL ORDER III: PUNISHMENT AND OTHER INCENTIVES

One important lesson from the research reported in the previous section why social order is fragile is that the only way a cooperator can avoid being “suckered” is to reduce his or her cooperation, thereby punishing everyone, even other cooperators. This raises the question whether targeted punishment (whereby group members can identify a free rider and punish

him or her) actually can solve the free rider problem and prevent the breakdown of cooperation. Mancur Olson, in a seminal analysis of the free rider problem in collective action, argued that "[O]nly a *separate and "selective" incentive* will stimulate a rational individual … to act in a group-oriented way". Olson further noted that selective incentives "can be either negative or positive, in that they can either coerce by punishing those who fail to bear an allocated share of the costs of the group action, or they can be positive inducements offered to those who act in the group interest" (Olson 1965, p.51, emphasis in original).

But who should apply these selective incentives? One answer is that in modern societies the legal system does the punishment. However, the state with its law enforcement institutions is a novel phenomenon on an evolutionary time scale. For a large part of human history, social order needed to be sustained without central institutions. And even in modern times, self-governance is often necessary in many important social dilemmas (Ostrom 1990).

One element of self-governance is informal sanctions as applied by other group members (Ostrom, Walker, and Gardner 1992; this is sometimes also called peer punishment). But the problem is that punishing is itself a public good: If Alice punishes a free rider who then subsequently contributes his or her share, Bill will benefit also, even if he has not punished (and thereby behaves as a "second-order free rider"). If Bill is a *homo economicus* he will certainly not punish if punishment is costly and has no personal benefit for him (which is likely in many situations), but if Alice is a *homo reciprocans* she might punish even if punishment is costly. The evidence on strong negative reciprocity, reported in Section 2, as well as the seminal studies by Yamagishi (1986) and Ostrom et al. (1992) suggest many people are indeed willing to punish free riders and the second-order public good problem is actually less of an issue. As reasoned above, free riders who fear punishment might have a selfish incentive to cooperate, and higher rates of cooperation should also convince conditional cooperators to keep cooperating.

Fehr and Gächter (2000) developed an experimental design to study punishment and cooperation in a sequence of ten one-shot (random group members - "Strangers") and fixed group ("Partners"; same group members) public good game – settings that correspond to different real-life interactions. The experiment proceeded as follows. Subjects first made their contributions to the public good, and then they entered a second stage, where they were informed about the individual contributions of all group members. Subjects could assign up

to ten punishment points to each individual group member. Punishment was costly for the punishing subject and each punishment point received reduced the punished subject's earnings from the first stage by ten percent.

The results support the *homo reciprocans* hypothesis that people are willing to punish free riders and that punishment increases cooperation. In both the Stranger and Partner conditions contributions increased over time – contrary to the *homo economicus* prediction. There is a substantial difference in cooperation rates between Partners and Strangers. Partners contributed about 85 percent of their maximal contribution and Strangers about 58 percent. By comparison, without punishment cooperation rates under Partners and Strangers were 38 and 19 percent, respectively. The fact that in the presence of punishment opportunities contributions even increased over time in a Strangers setting is particularly astonishing.

What explains the difference in cooperation between the Partners and Strangers condition? One likely channel is that at the cooperation stage within stable groups an interaction effect exists between the availability of punishment and strategic reciprocity (reciprocity that is also in the self-interest of a free rider due to the repeated nature of the interaction). A repeated interaction and punishment are complementary instruments to stimulate contributions. If only direct reciprocity is possible, cooperation collapses. If only punishment is possible but groups are formed randomly and hence direct reciprocity is not feasible, cooperation is stabilized at intermediate levels.

A theoretically interesting benchmark case of the Stranger condition is a situation where the likelihood of future interaction is zero, that is, groups interact only once in the same constellation. This situation is interesting, because evolutionary theories of cooperation (see Rand and Nowak 2013 for a succinct summary), predict no cooperation in this case. Therefore, Fehr and Gächter (2002) set up a so-called "perfect stranger" design where in each of the six repetitions all groups are composed of completely new members, and participants are aware of this. The results show again that cooperation increases over time when punishment is available.

The experiments of Fehr and Gächter (2000) and Fehr and Gächter (2002) also had a setting where subjects first played a condition with punishment and were then told that in a new condition the possibility of punishment would be removed. Again, the results show that punishment leads to high and stable cooperation rates. But when punishment is removed, cooperation collapses almost immediately and dwindles to low levels. This suggests that a

cooperative benchmark is not enough to support cooperation if not supported by the possibility of punishment.

While cooperation differs strongly between Partner, Stranger, and Perfect Stranger conditions, punishment patterns are qualitatively and even quantitatively similar across rounds: the more a group member deviates from the average contribution of his or her group members the higher is the punishment that he or she will receive. These observations are remarkable given that cooperation levels differ strongly between conditions. The fact that strong reciprocators punish even under Perfect Stranger conditions and that this punishment induces free riders to increase their contributions makes punishment altruistic: the punisher only bears the costs of punishment and because under Perfect Strangers the punisher will not meet the punished group member again the benefits of increased cooperation accrue solely to the future group members of the punished subject.

The experiments I have discussed so far force participants by way of experimental design into a condition where punishment is or is not available. What do people choose if they have a choice between being subjected to a condition where punishment is available and one where punishment is ruled out? Gürerk, Irlenbusch, and Rockenbach (2006) studied this question and got an interesting result. Initially, people opt for the no-punishment environment but soon they experience the problems of free riding. This experience changes their preferences and after a few more rounds the majority prefers an environment with punishment.

The proximate mechanisms behind altruistic punishment give an indication why punishment is not a second-order public good in practice. Punishment seems to be an impulse triggered by negative emotions and not much by forward-looking considerations (e.g., Casari and Luini 2012).

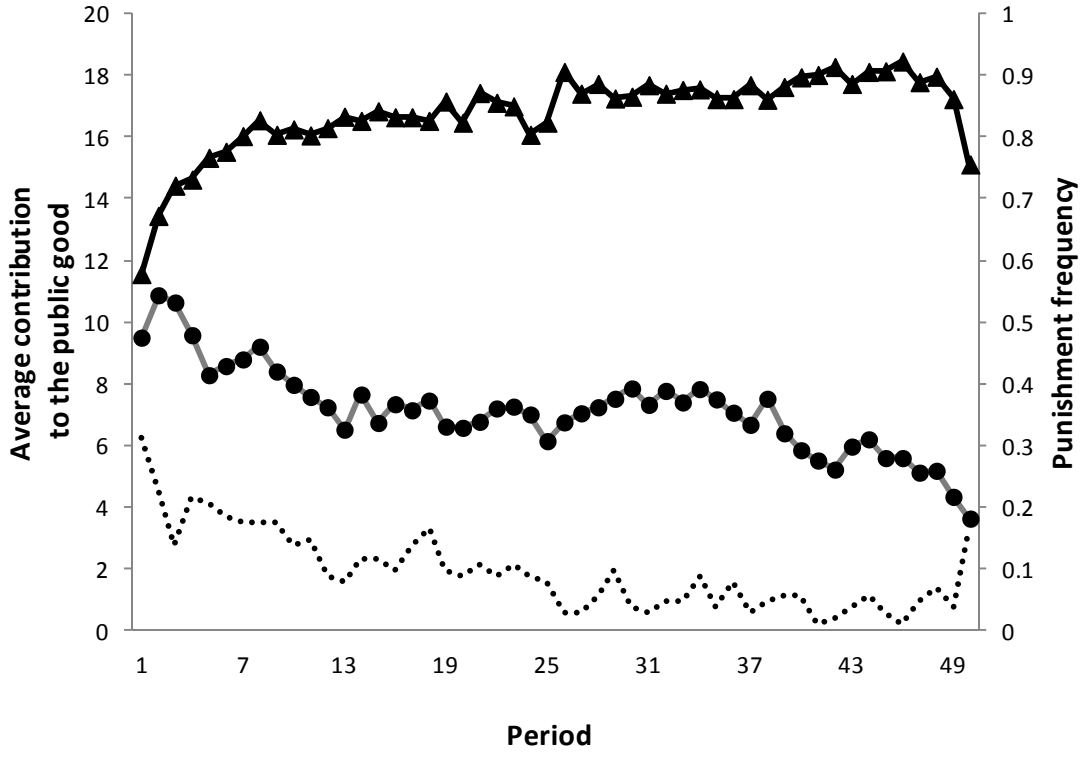
By now, there has been a lot of experimental and theoretical work on punishment and its effectiveness to stimulate cooperation. This literature is too voluminous to discuss here and I refer the interested reader to relevant surveys (Sigmund 2007; Gächter and Herrmann 2009; Balliet et al. 2011). I concentrate on five issues that are most relevant for my present purpose: the role of the severity of punishment and costs of punishment for the success of cooperation; punishment as a mere threat; imperfect observation and errors; institutionalized punishment; and incentives provided by rewards and reputation.

Severity and costs of punishment. The monitoring frequency and the severity of inflicted punishment matter for the effectiveness of punishment to stabilize (or increase) cooperation (Egas and Riedl 2008; Nikiforakis and Normann 2008). The more severe punishment is for the punished subject per unit of received punishment the higher are contributions. Although punishment is to a large extent non-strategic it follows cost-benefit considerations in the sense that punishment is less likely used the more costly it is for the punisher (e.g., Anderson and Putterman 2006). The fact that the level of cooperation corresponds to the severity of punishment suggests that low contributors respond strategically to the expected harm of punishment. If severe punishment is expected, free riders are deterred and cooperate. That is, although not pro-socially motivated, expected strong negative reciprocity can induce a selfishly motivated person to *behave* like a cooperator. Experiments by Shinada and Yamagishi (2007) also confirm the argument that increased cooperation by free riders through punishment strengthen the resolve of conditional cooperators to cooperate.

Punishment as a mere threat. One important characteristic feature of punishment is that it might not be used very often if people anticipate punishment and therefore try to avoid it through appropriate action. This is how law enforcement works in many instances. In the case of contributions to a public good, punishment is not necessary if people contribute at high levels and punishment might therefore simply act as a deterrent. This argument requires that punishment be a credible threat, that is, punishment indeed occurs if contributions are too low. If punishment is credible, then in equilibrium it will not happen very often. The existence of strong reciprocator suggests that some people are indeed willing to punish free riders, so punishment should be credible. After having received punishment free riders typically increase their contributions, so punishment has the desired behavioral effect. But can punishment also work as a mere threat?

To study the question whether punishment can also work as a mere threat, Gächter et al. (2008) extended the experiment to 50 periods. This should give plenty of time to establish punishment as a credible threat, and later on as a mere threat with very little actual punishment necessary to sustain high and stable contributions.

Figure 4: Punishment can stabilize social order through a threat of punishment alone.



Data source: Gächter et al. (2008). Own illustration. Triangles indicate the punishment condition; circles indicate the no punishment condition; and the dashed line indicates punishment frequency (measured on the right-hand vertical axis).

Figure 4 depicts cooperation with and without punishment. In the latter condition, cooperation is modest and slowly dwindling to low levels. In the condition with punishment cooperation approaches very high levels quickly. Consistent with the threat effect, punishment frequency is relatively high in the early phase of the experiment but approaches very low levels (less than 10 percent) in the second half of the experiment. Thus, punishment can exert its power as a mere threat effect, yet the threat has to be there. If punishment is impossible, cooperation breaks down.

Imperfect observation and errors. All experiments I have discussed so far assume that all contributions are perfectly observable and no errors occur. This is quite unrealistic and an important line of research investigates the consequences of imperfect observability and errors on punishment, cooperation, and overall efficiency of interactions. One way to model errors is to allow only for binary decisions: contribute or not (e.g., Ambrus and Greiner 2012). An

error occurs if a contribution is actually registered as a non-contribution with a certain probability. If people apply the legal principle that punishment should only be used if the true act is known, little punishment of non-contributions should occur. However, a typical finding is that people punish too much and falsely hit a contributor too often with the consequence that punishment is less effective in stimulating cooperation than under perfect error-free observability of contributions (Bornstein and Weisel 2010; Grechenig, Nicklisch, and Thöni 2010). See Grechenig et al. (2010) for a discussion of the relevance of these findings from a legal science point of view.

Institutionalized punishment. The evidence I have discussed so far is all based on peer punishment. These experiments reveal two things: people get angry about free riders (see Section 4) and this anger induces some people to punish free riders; that is, punishment reflects punitive sentiments. Given that punishment is expected, self-regarding people now have an incentive to cooperate. Modern lawful societies channel punitive sentiments into laws and a formal, institutionalized sanctioning system, which provide incentives to cooperate.

What matters from the point of view of a self-regarding individual is the expected cost of free riding. The presence of peer punishment might make cooperation worthwhile, but so can incentives provided by other mechanisms. For example, O'Gorman, Henrich, and Van Vugt (2009) and Baldassarri and Grossman (2011) studied centralized punishment by one group member and found it quite effective. Centralized punishment can even be effective if it is not deterrent (Engel 2013). Falkinger, Fehr, Gächter, and Winter-Ebmer (2000) showed that an exogenously given tax-subsidy mechanism induces people to cooperate in line with theoretical predictions about how the incentives should work. Another line of research, dating back to a seminal paper by Toshio Yamagishi (1986) showed that people are also willing to contribute to a "punishment fund" (think of funding law enforcement through people's taxes) to punish lowest contributors. Comparing (the evolution of) peer punishment and pool punishment has triggered theoretical investigations (Sigmund, De Silva, Traulsen, and Hauert 2010) and is also an important topic of experimental research (e.g., Traulsen, Röhl, and Milinski 2012; Zhang et al. 2013).

In the remainder of this section, I discuss briefly two mechanisms other than punishment that have also proved effective in supporting cooperation. The mechanisms I will consider are rewards; and indirect reciprocity and the role of a good reputation.

Rewards. Because punishment is successful in increasing cooperation (under perfect observability), an intuitive question is whether rewards can also sustain cooperation. Punishment, whenever it is used, has the disadvantage that it is costly for the punisher as well as for the punished person (i.e., punishment is inefficient because resources are destroyed). Rewards do not have this disadvantage. They might be costly too for the rewarding person, but if the benefits of the reward at least cover the costs, rewards are not inefficient.

Most experiments model rewards analogously to punishment: after group members have made their contributions, they are informed about each contribution made and can then allocate reward points to the target group member. One reward point costs 1 money unit and the rewarded group member then gets, depending on the experiment, one or more money units as an additional payment. The results suggest that this mechanism can also stimulate contributions, in particular if the rewarded individual receives more than what it costs to reward (Sefton, Shupp, and Walker 2007; Rand et al. 2009; Sutter, Haigner, and Kocher 2010). For example, in experiments comparable to Gächter et al. (2008) summarized in Figure 4, Rand et al. (2009) showed that achieved cooperation levels were as high as those under punishment.

It is important to notice that there is a fundamental asymmetry between punishments and rewards: rewards have to be used to be effective, whereas under punishment a credible threat can suffice (Figure 4). Thus, punishment can be very cheap whereas rewards will be costly. Moreover, in a context of law enforcement rewards are typically the exception and threats of punishment the norm.

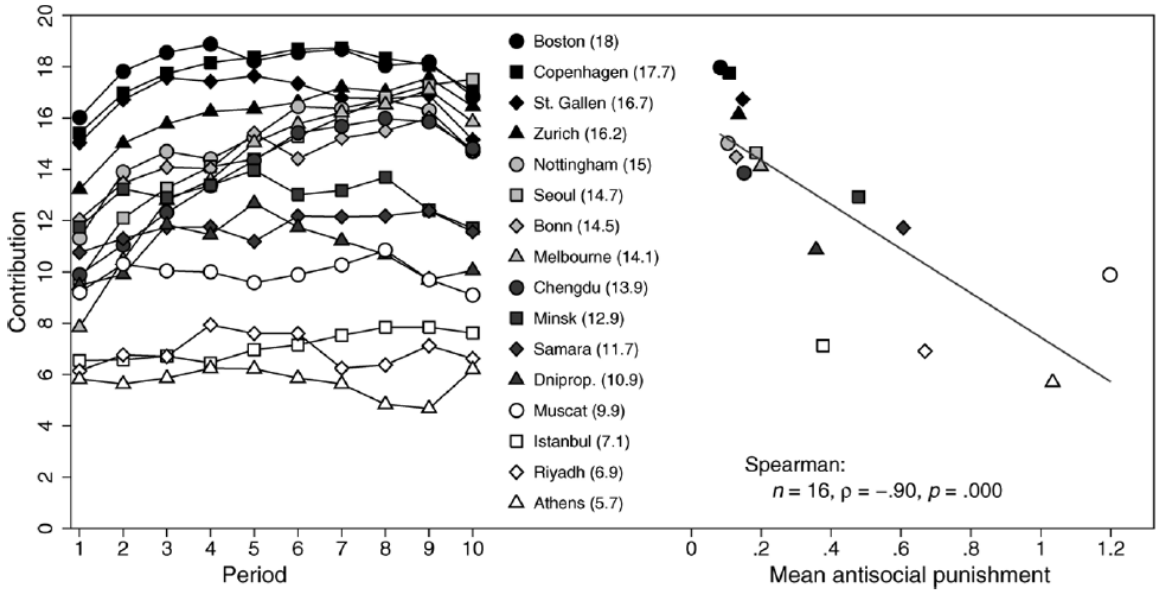
Indirect reciprocity and reputation. Humans keenly care about their reputation. Why? The mechanism of indirect reciprocity (Nowak and Sigmund 1998) provides an important likely channel. People not only help those who helped them (direct reciprocity) but might also help those who helped others. Thus, if one has a reputation of helping others one might receive more help as well and it pays to be a cooperator. Experimental evidence supports this theoretical argument (e.g., Milinski, Semmann, and Krambeck 2002). Relatedly, people's concerns to be held in good esteem can stimulate pro-social behavior (e.g., Ariely, Bracha, and Meier 2009). Evidence for the success of reputation-based incentives is not restricted to the lab. For example, a recent field experiment showed that a concern for good reputation can help in energy conservation, which is an important public good in the real world (Yoeli, Hoffman, Rand, and Nowak 2013).

6. RULE OF LAW AND SELF-GOVERNANCE OF SOCIAL DILEMMA PROBLEMS

The research I have presented so far has mostly been conducted in a few Western societies, such as the United States, Britain, and Switzerland. How representative are these societies when making claims or inferences about human nature? According to an influential study by Henrich, Heine, and Norenzayan (2010) there is substantial heterogeneity in human behavior across the many societies on this planet that make the Western societies look as outliers. Does this also hold for the behavioral patterns reported in this chapter?

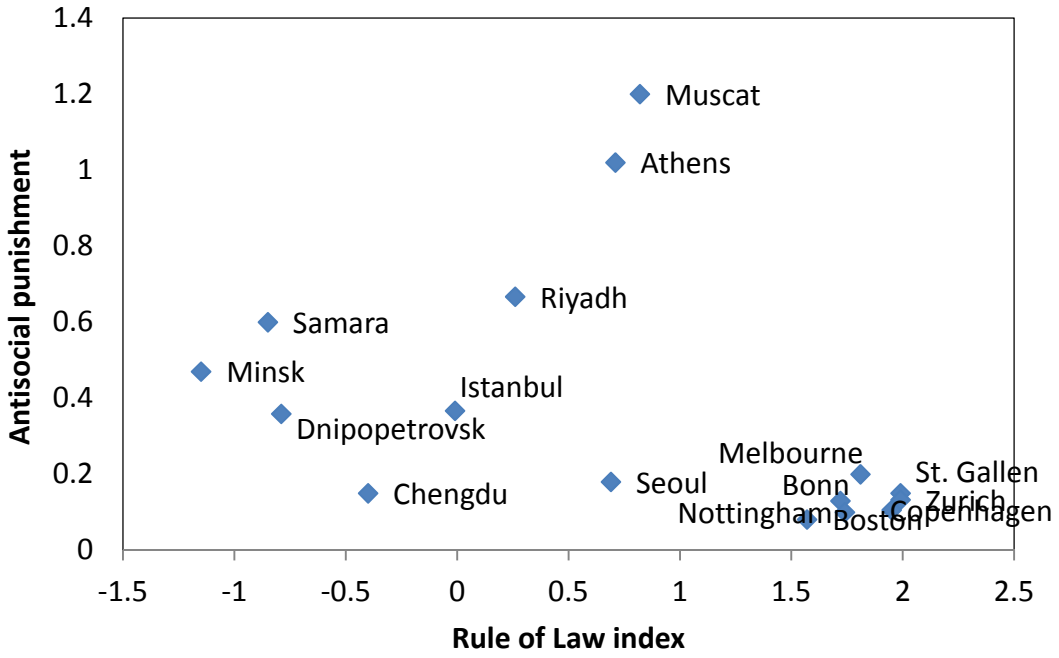
The existence of strong positive and negative reciprocity has been shown in many societies around the world (Henrich et al. 2005; Henrich et al. 2006). Herrmann et al. (2008) conducted a series of public goods experiments without and with punishment in fifteen quite different large-scale societies around the world (such as the United States, Turkey, China, Saudi Arabia, and England; see Figure 5). They uncovered three important findings relevant for the present topic. First, without punishment cooperation breaks down everywhere (Figure 3). Second, with punishment, it turns out that people punish free riders very similarly across the fifteen societies. In stark contrast, there is substantial cross-societal variation in antisocial punishment, that is, punishment of people who contributed to the public good by people who contributed less than the group member they punish. Third, there is a very large variation in cooperation levels achieved and, due to antisocial punishment, cooperation does not always raise contributions compared to the condition without punishment. Figure 5A illustrates the cooperation levels achieved and their relation to antisocial punishment.

Figure 5A: Cross-societal cooperation and antisocial punishment



Data source: Herrmann et al. (2008). The illustration is taken from Figure 8 in Gächter and Thöni (2011).

Figure 5B: The stronger is the Rule of Law the lower is antisocial punishment.



Data source: Herrmann et al. (2008); own illustration.

The results by Herrmann et al. (2008) provide us with an important caveat on the power of punishment to stimulate pro-social cooperation. Punishment only increases cooperation if it is targeted towards free riders exclusively; antisocial punishment is a huge impediment to successful cooperation. Relatedly, punishment can only stimulate cooperation if it does not trigger counter-punishment (e.g., Nikiforakis 2008).

The Herrmann et al. (2008) study reveals another relevant finding, namely that the severity of antisocial punishment in a society is linked to the Rule of Law in that society. The Rule of Law indicator is a governance indicator developed by the World Bank to measure how well private and government contracts can be enforced in courts, whether the legal system and police are perceived as being fair, how important the black market and organized crime are, etc. (see Herrmann et al. 2008 for details). Figure 5B illustrates how the Rule of Law is linked to antisocial punishment observed in a given society.

The results are quite striking. The Western societies all have a high Rule of Law index value and there is also very little antisocial punishment observed in these societies. The variation in antisocial punishment increases substantially once the Rule of Law index falls below 1 (the theoretical range is between -2.5 and +2.5).

The significance of this finding is twofold. First, the fact that experimentally measured behavior is correlated to societal measures suggests that the societal background has an influence on behavior. The studies by Henrich et al. (2005), Henrich et al. (2006) and Henrich, Ensminger, et al. (2010) suggest such an influence based on the organization of the small-scale societies where they conducted their research. The Herrmann et al. (2008) findings show that societal background also matters for developed, large-scale societies. Second, and more importantly for present purposes, the negative correlation of antisocial punishment and the quality of the Rule of Law in a society suggests that a high quality law enforcement system (which can be interpreted as a high degree of institutionalized cooperation) will also limit antisocial punishment and thereby an important inhibitor of voluntary cooperation. Good institutions make for good self-governance of people who manage to cooperate with one another and who limit punishment to those who fail to cooperate.

7. SUMMARY AND CONCLUDING REMARKS

In this chapter I have provided evidence from two decades of behavioral economics research that, rather than being selfish as is assumed in the *homo economicus* paradigm, many people are strong reciprocators, who punish wrongdoing and reward kind acts. However, a sizeable minority of people is best characterized as selfish. My main focus has been on determinants of social order which I have construed as a social dilemma where individual incentives are not aligned with collective benefits.

I have argued that from the perspective of the behavioral science of cooperation, and in particular strong reciprocity, social order has three important determinants: the strength of internalized norms of pro-social behavior, the behavior of other people, and the threat of punishment or the presence of other incentives to curb selfishness. Looking at the many results in synthesis suggests the following big picture: many people are motivated by character virtues such as honesty and trustworthiness; they think that free riding is morally blameworthy; they feel guilty if it turns out that others contributed more to the public good than them; they are angry at the free riders; and they experience some warm glow by contributing to the public good. However, all research shows that people are also very strongly looking at the behavior of others to determine their behavior. Since a sizeable number of people are free riders and even many conditional cooperators have a selfish bias, cooperation in randomly assembled groups is inherently fragile. Cooperation can only be sustained under the strong requirement that only highly cooperatively inclined people are matched and able to exclude free rider types. Under more realistic conditions, stable pro-social cooperation requires some incentives, most notably punishment, where often a credible threat suffices to keep free riding at bay.

Notice that the three determinants of social order are also linked. If norms are strong and induce many people to cooperate then the psychology of conditional cooperation will induce many people to cooperate as well. However, because a sizeable minority of people is not motivated by normative considerations but only by own gain, norms appear a rather weak determinant of social order because conditional cooperators will only cooperate if others do as well. In other words, the psychology of conditional cooperation appears to be the stronger behavioral force than normative considerations and, as a consequence, cooperation will be fragile. This conclusion follows from three separate observations I recorded in this chapter: (i) character virtues and normative considerations including feelings of guilt if others behave

more cooperatively matter for many people (Figure 1); (ii) conditional cooperation is an important motivation for the average person (Figure 3) and (iii) cooperation nevertheless almost inevitably breaks down if not backed up by incentives (Figure 2). Punishment (or other incentives) have the dual advantage that they induce the free rider types to cooperate and thereby convince the conditional cooperators to maintain their cooperation.

I conclude this chapter with some remarks on future research. Of the three determinants of social order the first determinant (the role of norms, moral judgments and emotions such as guilt for cooperation) is the least well understood determinant of cooperation. More research is necessary to understand people's normative consideration and to what extent this influences their behavior. With regard to the second determinant (conditional cooperation) an important open question is gaining a complete picture of proximate mechanisms that determine conditional cooperation including gauging their relative importance. The third determinant (punishment and other incentives) is the best understood determinant. Open questions are finding explanations for antisocial punishment and how antisocial punishment is causally related to the Rule of Law (Figure 5). A further under-researched topic is the role of institutional punishment for successful cooperation, in particular in comparison with peer punishment and when considering the role of errors and imperfect observability. Finally, an important topic for future research is to understand how exactly the three determinants are linked and how the three determinants work in naturally occurring settings.

REFERENCES

Ambrus, A., and B. Greiner. 2012. Imperfect public monitoring with costly punishment: An experimental study. *American Economic Review* 102:3317-32.

Anderson, C. M., and L. Putterman. 2006. Do non-strategic sanctions obey the law of demand? The demand for punishment in the voluntary contribution mechanism. *Games and Economic Behavior* 54:1-24.

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

Ariely, D., A. Bracha, and S. Meier. 2009. Doing good or doing well? Image motivation and monetary incentives in behaving prosocially. *American Economic Review* 99:544-555.

Axelrod, R. 1984. *The evolution of cooperation*. New York: Basic Books.

Baldassarri, D., and G. Grossman. 2011. Centralized sanctioning and legitimate authority promote cooperation in humans. *Proceedings of the National Academy of Sciences* 108:11023-11027.

Balliet, D., L. B. Mulder, and P. A. M. Van Lange. 2011. Reward, punishment, and cooperation: A meta-analysis. *Psychological Bulletin* 137:594-615.

Balliet, D., C. Parks, and J. Joireman. 2009. Social value orientation and cooperation in social dilemmas: A meta-analysis. *Group Processes & Intergroup Relations* 12:533-547.

Baron-Cohen, S. 2011. *Zero degrees of empathy. A new theory of human cruelty*. London: Allen Lane.

Berg, J., J. Dickhaut, and K. McCabe. 1995. Trust, reciprocity, and social history. *Games and Economic Behavior* 10:122-142.

Bolton, G. E., and A. Ockenfels. 2000. ERC: A theory of equity, reciprocity, and competition. *American Economic Review* 90:166-93.

Bornstein, G., and O. Weisel. 2010. Punishment, cooperation, and cheater detection in "noisy" social exchange. *Games* 1:18-33.

Bowles, S., and H. Gintis. 2011. *A cooperative species: Human reciprocity and its evolution*. Princeton: Princeton University Press.

Boyd, R., and P. J. Richerson. 1988. The evolution of reciprocity in sizable groups. *Journal of Theoretical Biology* 132:337-356.

Brekke, K. A., G. Kipperberg, and K. Nyborg. 2010. Social interaction in responsibility ascription: The case of household recycling. *Land Economics* 86:766-784.

Buckholtz, J. W., and R. Marois. 2012. The roots of modern justice: Cognitive and neural foundations of social norms and their enforcement. *Nature Neuroscience* 15:655-661.

Carpenter, J. 2004. When in Rome: Conformity and the provision of public goods. *Journal of Socio-Economics* 33:395-408.

Casari, M., and L. Luini. 2012. Peer punishment in teams: Expressive or instrumental choice? *Experimental Economics* 15:241-259.

Charness, G., and P. Kuhn. 2011. Lab labor: What can labor economists learn from the lab? In *Handbook of labor economics*, O. Ashenfelter, and D. Card. Amsterdam: Elsevier, 229-330.

Charness, G., and M. Rabin. 2002. Understanding social preferences with simple tests. *Quarterly Journal of Economics* 117:817-69.

Chaudhuri, A. 2011. Sustaining cooperation in laboratory public goods experiments: A selective survey of the literature. *Experimental Economics* 14:47-83.

Chaudhuri, A., and T. Paichayontvijit. 2006. Conditional cooperation and voluntary contributions to a public good. *Economics Bulletin* 3:1-15.

Cooter, R. 2000. Do good laws make good citizens? An economic analysis of internalized norms. *Virginia Law Review* 86:1577-1601.

Cubitt, R., M. Drouvelis, and S. Gächter. 2011a. Framing and free riding: Emotional responses and punishment in social dilemma games. *Experimental Economics* 14:254–272.

Cubitt, R., M. Drouvelis, S. Gächter, and R. Kabalin. 2011b. Moral judgments in social dilemmas: How bad is free riding? *Journal of Public Economics* 95:253-264.

Dufwenberg, M., S. Gächter, and H. Hennig-Schmidt. 2011. The framing of games and the psychology of play. *Games and Economic Behavior* 73:459-478.

Dufwenberg, M., and G. Kirchsteiger. 2004. A theory of sequential reciprocity. *Games and Economic Behavior* 47:268-298.
Eckel, C., and P. Grossman. 1996. Altruism in anonymous dictator games. *Games and Economic Behavior* 16:181-191.
Egas, M., and A. Riedl. 2008. The economics of altruistic punishment and the maintenance of cooperation. *Proceedings of the Royal Society B - Biological Sciences* 275:871-878.
Ehrhart, K.-M., and C. Keser. 1999. Mobility and cooperation: On the run. *Working Paper No. 99s-24, CIRANO Montreal*.
Ellickson, R. 1991. *Order without law. How neighbors settle disputes*. Cambridge: Harvard University Press
Engel, C. 2011. Dictator games: A meta study. *Experimental Economics* 14:583-610.
Engel, C. 2013. Deterrence by imperfect sanctions - a public good experiment. *Preprints of the Max Planck Institute for Research on Collective Goods Bonn 2013/9*.
Engelmann, D., and M. Strobel. 2004. Inequality aversion, efficiency, and maximin preferences in simple distribution experiments. *American Economic Review* 94:857-69.
Falk, A., E. Fehr, and U. Fischbacher. 2003. On the nature of fair behavior. *Economic Inquiry* 41:20-26.
Falk, A., E. Fehr, and U. Fischbacher. 2005. Driving forces behind informal sanctions. *Econometrica* 73:2017-2030.
Falk, A., and U. Fischbacher. 2006. A theory of reciprocity. *Games and Economic Behavior* 54:293-315.
Falk, A., and J. Heckman. 2009. Lab experiments are a major source of knowledge in the social sciences. *Science* 326:535-538.
Falkinger, J., E. Fehr, S. Gächter, and R. Winter-Ebmer. 2000. A simple mechanism for the efficient provision of public goods: Experimental evidence. *American Economic Review* 90:247-264.
Fehr, E., and U. Fischbacher. 2004. Social norms and human cooperation. *TRENDS in Cognitive Sciences* 8:185-190.
Fehr, E., U. Fischbacher, and S. Gächter. 2002. Strong reciprocity, human cooperation, and the enforcement of social norms. *Human Nature* 13:1-25.
Fehr, E., and S. Gächter. 2000. Cooperation and punishment in public goods experiments. *American Economic Review* 90:980-994.
Fehr, E., and S. Gächter. 2002. Altruistic punishment in humans. *Nature* 415:137-140.
Fehr, E., G. Kirchsteiger, and A. Riedl. 1993. Does fairness prevent market clearing? An experimental investigation. *Quarterly Journal of Economics* 108:437-459.
Fehr, E., and K. M. Schmidt. 1999. A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics* 114:817-68.
Fehr, E., and K. M. Schmidt. 2006. The economics of fairness, reciprocity and altruism - experimental evidence and new theories. In *Handbook of the economics of giving, altruism and reciprocity*, S.-C. Kolm, and J. M. Ythier. Amsterdam: Elsevier B.V., 615-691.
Fischbacher, U., and S. Gächter. 2010. Social preferences, beliefs, and the dynamics of free riding in public good experiments. *American Economic Review* 100:541–556.
Fischbacher, U., S. Gächter, and E. Fehr. 2001. Are people conditionally cooperative? Evidence from a public goods experiment. *Economics Letters* 71:397-404.
Fischbacher, U., S. Gächter, and S. Quercia. 2012. The behavioral validity of the strategy method in public good experiments. *Journal of Economic Psychology* 33:897-913.
Forsythe, R., J. Horowitz, N. E. Savin, and M. Sefton. 1994. Fairness in simple bargaining experiments. *Games and Economic Behavior* 6:347-69.
Frey, B. S., and S. Meier. 2004. Social comparisons and pro-social behavior. Testing 'conditional cooperation' in a field experiment. *American Economic Review* 94:1717-1722.
Frey, B. S., and B. Torgler. 2007. Tax morale and conditional cooperation. *Journal of Comparative Economics* 35:136-159.
Friedman, D., and S. Sunder. 1994. *Experimental methods. A primer for economists*. Cambridge: Cambridge University Press.

Gächter, S., and B. Herrmann. 2009. Reciprocity, culture, and human cooperation: Previous insights and a new cross-cultural experiment. *Philosophical Transactions of the Royal Society B – Biological Sciences* 364:791-806.
Gächter, S., E. Renner, and M. Sefton. 2008. The long-run benefits of punishment. *Science* 322:1510.
Gächter, S., and C. Thöni. 2005. Social learning and voluntary cooperation among like-minded people. *Journal of the European Economic Association* 3:303-314.
Gächter, S., and C. Thöni. 2011. Micromotives, microstructure and macrobehavior: The case of voluntary cooperation. *Journal of Mathematical Sociology* 35:26-65.
Galbiati, R., and P. Vertova. 2008. Obligations and cooperative behaviour in public good games. *Games and Economic Behavior* 64:146-170.
Gintis, H. 2000. Strong reciprocity and human sociality. *Journal of Theoretical Biology* 206:169-179.
Gintis, H., S. Bowles, R. Boyd, and E. Fehr, Eds. 2005. *Moral sentiments and material interests. The foundations of cooperation in economic life*. Cambridge: MIT Press.
Glimcher, P. W., C. F. Camerer, E. Fehr, and R. A. Poldrack, Eds. 2009. *Neuroeconomics. Decision making and the brain*. Amsterdam: Elsevier.
Gneezy, U. 2005. Deception: The role of consequences. *American Economic Review* 95 1:384-94.
Grechenig, C., A. Nicklisch, and C. Thöni. 2010. Punishment despite reasonable doubt - a public goods experiment with uncertainty over contributions. *Journal of Empirical Legal Studies* 7:847-867.
Grujić, J., B. Eke, A. Cabrales, J. A. Cuesta, and A. Sánchez. 2012. Three is a crowd in iterated prisoner's dilemmas: Experimental evidence on reciprocal behavior. *Scientific Reports* 2:10.1038/srep00638.
Gürerk, Ö., B. Irlenbusch, and B. Rockenbach. 2006. The competitive advantage of sanctioning institutions. *Science* 312:108-111.
Güth, W., R. Schmittberger, and B. Schwarze. 1982. An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior and Organization* 3:367-88.
Haidt, J. 2003. The moral emotions. In *Handbook of affective sciences*, R. J. Davidson, K. R. Scherer, and H. H. Goldsmith. Oxford: Oxford University Press, 852–870.
Henrich, J., R. Boyd, S. Bowles, C. F. Camerer, E. Fehr, and H. Gintis. 2004. *Foundations of human sociality. Economic experiments and ethnographic evidence from fifteen small-scale societies*. Oxford: Oxford University Press.
Henrich, J., R. Boyd, S. Bowles, C. F. Camerer, E. Fehr, H. Gintis, R. McElreath, M. Alvard, A. Barr, J. Ensminger, N. Henrich, K. Hill, F. Gil-White, M. Gurven, F. W. Marlowe, J. Q. Patton, and D. Tracer. 2005. "Economic man" in cross-cultural perspective: Behavioral experiments in 15 small-scale societies. *Behavioral and Brain Sciences* 28:795-855.
Henrich, J., J. Ensminger, R. McElreath, A. Barr, C. Barrett, A. Bolyanatz, J. C. Cardenas, M. Gurven, E. Gwako, N. Henrich, C. Lesorogol, F. Marlowe, D. Tracer, and J. Ziker. 2010. Markets, religion, community size, and the evolution of fairness and punishment. *Science* 327:1480-1484.
Henrich, J., S. J. Heine, and A. Norenzayan. 2010. The weirdest people in the world? *Behavioral and Brain Sciences* 33:61-83.
Henrich, J., R. McElreath, A. Barr, J. Ensminger, C. Barrett, A. Bolyanatz, J.-C. Cardenas, M. Gurven, E. Gwako, N. Henrich, C. Lesorogol, F. Marlowe, T. David, and J. Ziker. 2006. Costly punishment across human societies. *Science* 312:1767-1770.
Herrmann, B., and C. Thöni. 2009. Measuring conditional cooperation: A replication study in russia. *Experimental Economics* 12:87-92.
Herrmann, B., C. Thöni, and S. Gächter. 2008. Antisocial punishment across societies. *Science* 319:1362-1367.
Hume, D. 1987 [1777]. *Essays: Moral, political and literary*. Indianapolis: Liberty Fund Inc.
Johnson, N. D., and A. A. Mislin. 2011. Trust games: A meta-analysis. *Journal of Economic Psychology* 32:865-889.
Kahan, D. 2003. The logic of reciprocity: Trust, collective action, and law. *Michigan Law Review* 102:71-103.
Kahneman, D., and A. Tversky, Eds. 2000. *Choices, values, and frames*. Cambridge: Cambridge University Press.

Kirchgässner, G. 2008. *Homo oeconomicus: The economic model of behaviour and its applications in economics and other social sciences*. New York: Springer.
Kirchler, E. 2007. *The economic psychology of tax behaviour*. Cambridge: Cambridge University Press.
Kocher, M. G., T. Cherry, S. Kroll, R. J. Netzer, and M. Sutter. 2008. Conditional cooperation on three continents. *Economics Letters* 101:175-178.
Lewinsohn-Zamir, D. 1998. Consumer preferences, citizen preferences, and the provision of public goods. *Yale Law Journal* 108:377-406.
López-Pérez, R., and E. Spiegelman. 2013. Why do people tell the truth? Experimental evidence for pure lie aversion. *Experimental Economics* 16:233-247.
Martinsson, P., N. Pham-Khanh, and C. Villegas-Palacio. 2013. Conditional cooperation and disclosure in developing countries. *Journal of Economic Psychology* 34:148-155.
Milinski, M., D. Semmann, and H. J. Krambeck. 2002. Reputation helps solve the 'tragedy of the commons'. *Nature* 415:424-426.
Nikiforakis, N. 2008. Punishment and counter-punishment in public good games: Can we really govern ourselves? *Journal of Public Economics* 92:91-112.
Nikiforakis, N., and H. Normann. 2008. A comparative statics analysis of punishment in public goods experiments. *Experimental Economics* 11:358-369.
Nowak, M. A., and K. Sigmund. 1998. Evolution of indirect reciprocity by image scoring. *Nature* 393:573-577.
O'Gorman, R., J. Henrich, and M. Van Vugt. 2009. Constraining free riding in public goods games: Designated solitary punishers can sustain human cooperation. *Proceedings of the Royal Society B-Biological Sciences* 276:323-329.
Olson, M. 1965. *The logic of collective action*. Cambridge Harvard University Press.
Oosterbeek, H., R. Sloof, and G. van de Kuilen. 2004. Cultural differences in ultimatum game experiments: Evidence from a meta-analysis. *Experimental Economics* 7:171-188.
Ostrom, E. 1990. *Governing the commons. The evolution of institutions for collective action*. Cambridge: Cambridge University Press.
Ostrom, E., J. M. Walker, and R. Gardner. 1992. Covenants with and without a sword - self-governance is possible. *American Political Science Review* 86:404-417.
Posner, E. A. 2000. *Law and social norms*. Cambridge: Harvard University Press.
Pruckner, G. J., and R. Sausgruber. 2013. Honesty on the streets: A field study on newspaper purchasing. *Journal of the European Economic Association* 11:661-679.
Rabin, M. 1993. Incorporating fairness into game-theory and economics. *American Economic Review* 83:1281-1302.
Rand, D. G., A. Dreber, T. Ellingsen, D. Fudenberg, and M. A. Nowak. 2009. Positive interactions promote public cooperation. *Science* 325:1272-1275.
Rand, D. G., and M. A. Nowak. 2013. Human cooperation. *Trends in Cognitive Sciences* 17:413-425.
Rapoport, A., and A. M. Chammah. 1965. *Prisoners' dilemma. A study in conflict and cooperation*. Ann Arbor: The University of Michigan Press.
Riker, W. H., and P. C. Ordeshook. 1968. A theory of the calculus of voting. *The American Political Science Review* 62:25-42.
Rustagi, D., S. Engel, and M. Kosfeld. 2010. Conditional cooperation and costly monitoring explain success in forest commons management. *Science* 330:961-965.
Schwartz, S. H. 1977. Normative influences on altruism. In *Advances in experimental social psychology*, B. Leonard. Academic Press, 221-279.
Sefton, M., R. Shupp, and J. M. Walker. 2007. The effect of rewards and sanctions in provision of public goods. *Economic Inquiry* 45:671–690.
Shang, J., and R. Croson. 2009. A field experiment in charitable contribution: The impact of social information on the voluntary provision of public goods. *Economic Journal* 119:1422-1439.
Shinada, M., and T. Yamagishi. 2007. Punishing free riders: Direct and indirect promotion of cooperation. *Evolution and Human Behavior* 28:330-339.
Sigmund, K. 2007. Punish or perish? Retaliation and collaboration among humans. *TRENDS in Ecology and Evolution* 22:593-600.

Sigmund, K., H. De Silva, A. Traulsen, and C. Hauert. 2010. Social learning promotes institutions for governing the commons. *Nature* 466:861-863.
Stigler, G. J. 1981. Economics or ethics? In *Tanner lectures on human values*, S. McMurrin. Cambridge: Cambridge University Press,
Sunstein, C., Ed. 2000. *Behavioral law and economics*. Cambridge: Cambridge University Press.
Sutter, M., S. Haigner, and M. G. Kocher. 2010. Choosing the carrot or the stick? Endogenous institutional choice in social dilemma situations. *Review of Economic Studies* 77:1540-1566.
Thöni, C., J.-R. Tyran, and E. Wengström. 2012. Microfoundations of social capital. *Journal of Public Economics* 96:635-643.
Traulsen, A., T. Röhl, and M. Milinski. 2012. An economic experiment reveals that humans prefer pool punishment to maintain the commons. *Proceedings of the Royal Society B: Biological Sciences* 279:3716-3721.
Traxler, C. 2010. Social norms and conditional cooperative taxpayers. *European Journal of Political Economy* 26:89-103.
Van Lange, P. A. M., D. Balliet, C. D. Parks, and M. Van Vugt. 2014. *Social dilemmas. The psychology of human cooperation*. Oxford: Oxford University Press.
Volk, S., C. Thöni, and W. Ruigrok. 2012. Temporal stability and psychological foundations of cooperation preferences. *Journal of Economic Behavior & Organization* 81:664-676.
Wilkinson, N., and M. Klaes. 2012. *An introduction to behavioral economics, 2nd edition*. Basingstoke: Palgrave Macmillan.
Yamagishi, T. 1986. The provision of a sanctioning system as a public good. *Journal of Personality and Social Psychology* 51:110-116.
Yamagishi, T., N. Mifune, Y. Li, M. Shinada, H. Hashimoto, Y. Horita, A. Miura, K. Inukai, S. Tanida, T. Kiyonari, H. Takagishi, and D. Simunovic. 2013. Is behavioral pro-sociality game-specific? Pro-social preference and expectations of pro-sociality. *Organizational Behavior and Human Decision Processes* 120:260-271.
Yoeli, E., M. Hoffman, D. G. Rand, and M. A. Nowak. 2013. Powering up with indirect reciprocity in a large-scale field experiment. *Proceedings of the National Academy of Sciences* 110:10424-10429.
Zhang, B., C. Li, H. Silva, P. Bednarik, and K. Sigmund. 2013. The evolution of sanctioning institutions: An experimental approach to the social contract. *Experimental Economics* 1-19.