

Covenants before the swords: The limits to efficient cooperation in heterogeneous groups*

Christian Koch,^a Nikos Nikiforakis,^b and Charles N. Noussair^c

This version: May 21, 2020

Abstract

When agents derive heterogeneous benefits from cooperation, a tension often arises between efficiency and equality that can impede their ability to cooperate efficiently. We design a lab experiment, in which we investigate the efficacy of communication and punishment, separately and jointly, to promote cooperation in such an environment. Our results reveal that communication allows most groups to establish *covenants*, i.e., agreements about the profile of individual contributions, while the threat of punishment (the ‘sword’) discourages deviations from the covenants. Most covenants, however, reflect a concern for equality. As a result, cooperation levels and earnings fall substantially below the maximum possible. The timing of communication is also critical: covenants reduce the use of sanctions dramatically when communication precedes punishment opportunities but when punishment precedes communication opportunities, a history of sanctioning emerges which reduces the probability that groups establish covenants subsequently. Our findings illustrate not only the benefits of early communication, but also some limits to self-governance in heterogeneous groups.

Keywords: communication, punishment, cooperation, normative conflict, heterogeneity

JEL Codes: C92, H41, D74

^a *Corresponding Author:* Department of Economics, University of Vienna, Oskar-Morgenstern Platz 1, 1090 Vienna. chris.koch@univie.ac.at

^b Division of Social Science, New York University Abu Dhabi, P.O. Box 129188, Abu Dhabi, United Arab Emirates. nikos.nikiforakis@nyu.edu

^c Department of Economics, Eller College of Management, University of Arizona, Tucson, Arizona 85721, USA, cnoussair@email.arizona.edu

* We would like to thank Tim Cason, Dirk Engelmann, Francoise Forges, Arno Riedl, seminar participants at the HU/WZB/DIW Berlin, and attendees at the 2016 Maastricht Behavioral and Experimental Symposium (M-BEES), the 2016 ESA World Meeting in Jerusalem, the 2017 International Symposium on Experimental Economics (iSee) in Abu Dhabi, the 2017 ESA World Meeting in San Diego, and the 2018 ESA World Meeting in Berlin. Declaration of interest: none.

1. Introduction

The ability of groups to cooperate in the absence of central authority has been a topic of continuing interest among economists. Several studies have shown that individuals can achieve high – even maximal – levels of cooperation if they are provided with the means to discipline free riders (e.g., Fehr and Gächter 2000, 2002; Masclet et al. 2003). However, most of these studies have focused on a special case in which cooperation yields the same benefits to all agents so that everyone gains equally from solving the free-rider problem. The emphasis on homogeneous agents is natural when one is interested in studying, in isolation, how the tension between private and public interest – the defining feature of all social dilemmas – affects the ability of groups to cooperate. In daily life, however, agents often differ in the benefits they derive from cooperation; countries differ in the benefits they gain from alliances and treaties; firms differ in the benefits they obtain from forming cartels due to differences in capacity and cost structure; and individuals in working groups differ in motivation and career plans. It is therefore important to understand whether such differences between agents can be an obstacle to cooperation.

There are good reasons why this could be the case. Heterogeneity adds another layer of complexity to social dilemmas. When agents obtain different benefits from cooperation, as we will see below, they can face a trade-off between equality and efficiency: higher levels of cooperation can increase both group earnings (a measure of efficiency) as well as inequality. This *normative conflict* implies that agents must agree, not only about the need to cooperate, but also about the appropriate level of cooperation of each party involved. Failure to do so could lead cooperation to unravel. Overcoming normative conflict is, therefore, necessary for efficient cooperation, but it is not sufficient. Even if agents can resolve the normative conflict and agree on individual contributions, concerns for equality among group members can still limit the efficiency gains obtained from cooperation. The question that arises, therefore, is whether agents who derive differing benefits from cooperation should be expected to efficiently cooperate with each other (i.e., maximize the gains from cooperation) when this entails a certain level of inequality. If cooperation levels fall short of the maximum, it is important to understand the reasons this happens as this will allow us to identify policies that can promote welfare.

Can groups whose members derive heterogeneous benefits cooperate efficiently? Recent studies suggest that the ability to sanction free riders is less effective at promoting cooperation in heterogeneous groups (Reuben and Riedl 2013), especially when punishment can escalate and lead to feuds (Nikiforakis *et al.* 2012). While these studies provide interesting insights about the difficulties normative conflict poses for cooperation, it seems reasonable to assume that the inability of individuals to communicate with each other in these studies exaggerates the problem

posed by normative conflict. Similarly, Gangadharan et al. (2017) show that communication is less effective at promoting cooperation in heterogeneous groups, but the inability to sanction other group members could, again, cause us to overestimate the problem of normative conflict. For example, the lower levels of cooperation could be attributed to the greater propensity of individuals to violate agreements made during the communication phase or to push for efficient cooperation.

In this paper, we present evidence from a new laboratory experiment investigating the effect of communication and punishment on cooperation, separately and jointly. Using a simple theoretical framework to understand the implications of normative conflict, we show that efficient cooperation is unlikely to emerge when *either* communication or punishment is not possible – as is the case in the studies mentioned above. We introduce a conceptual distinction between *ex-ante* and *ex-post normative conflict*. The former is defined as the uncertainty individuals face about the extent to which other agents may care about the different normative criteria, which can be greatly alleviated through communication. Ex-post normative conflict arises when proponents of different normative criteria coexist in a group. Punishment can help mitigate the latter problem by providing a coercive force that helps group members find a compromise. Thus, when both instruments act in tandem, efficient cooperation is attainable even if agents hold different normative views, unless there exist agents who care too strongly about equality. Such agents will not only refuse to cooperate fully, but may also, in some instances, engage in costly feuding, i.e., cycles of reciprocal sanctions, against those championing different normative rules.

To better understand the limits of our findings, we explore how critical the *timing* of communication is for cooperation – a topic of significance in international relations (e.g., Regan and Stam 2000). If communication helps agents agree on what constitutes an appropriate level of cooperation, then its timing could be important. Group members that – potentially due to ex-ante normative conflict – have been caught up in a feud may be less willing to cooperate with each other and in such instances communication may be unable to overcome a history of sanctioning.¹ In order to explore this issue, communication is introduced at the onset of players' interaction in some treatments. In the remaining treatments, it is introduced midway through the interaction.

Our setting is a variation of the public-good game – a workhorse paradigm in the social sciences. As in Nikiforakis *et al.* (2012), our environment features heterogeneous preferences along with rich punishment opportunities. Unlike in most other studies about punishment, agents cannot only sanction others, but also counter-sanction, and even engage in retaliatory cycles of sanctioning

¹ Throughout the paper, we will try to distinguish the mere availability of punishment opportunities from their actual application by referring to the former using the term (*threat of*) *punishment*, and referring to the latter with the term *sanctioning* or by making explicit that we mean the actual application of punishment.

(feuds). Agents differ in the benefits they derive from cooperation, so that higher efficiency typically leads to greater inequality. Importantly, benefits are determined by individuals' performance in a real-effort task so that concerns for efficiency and equity are aligned against those for equality. In our setting, the efficient outcome is an equitable outcome in the sense that it respects the previously earned difference in benefits agents receive from cooperation. This also implies that equality should be *less* attractive as a normative rule (e.g., Cappelen *et al.* 2007, Gee *et al.* 2017) than it would be in the absence of the real effort task. In other words, we deliberately chose an environment in which we have reasons to believe equality concerns should play a relatively limited role and normative conflict should be mostly (if not only) *ex-ante* rather than *ex-post*. Therefore, the conflict should be relatively easy to overcome through peer-to-peer communication. We argue that failure to observe efficient cooperation in our setting indicates the significance of equality concerns and, by extension, the limits to efficient cooperation in heterogeneous groups.

Although concerns for equality appear to play an important role in zero-sum games, their relative importance in positive-sum decisions such as those involved in social dilemmas is unclear. Several studies have found that concerns for efficiency (measured as the sum of group earnings) dominate those for equality in payoffs (e.g., Balafoutas *et al.* 2012, Cabrales *et al.* 2010, Charness and Rabin 2002, Engelmann and Strobel 2004, Faravelli *et al.* 2013, Fisman *et al.* 2007). Bland and Nikiforakis (2015), furthermore, find that even individuals revealed to care about equality in zero-sum games tend to ignore inequalities when trying to coordinate.² This evidence suggests that *ex-post* normative conflict may not be a major obstacle to efficient cooperation, and that previous studies may have been capturing the effect of *ex-ante* normative conflict.³ This is especially true in an environment such as ours, where individuals concerned for equity and efficiency would favor the same contribution rule. In line with this, Gangadharan *et al.* (2017) find no evidence of individuals using monetary transfers to reduce inequality in heterogeneous groups. Taken together, this evidence implies that efficient cooperation may indeed be attainable in heterogeneous groups if their members are provided with the appropriate means.

² This observation is interesting for our purposes. Fehr and Schmidt (1999, Proposition IV) show that peer punishment can transform a social dilemma to a coordination game with multiple (Pareto-ranked) equilibria. In line with this, Nikiforakis (2010) shows that subtle changes in the institutional environment that do not affect monetary incentives or the use of punishment lead to substantially different cooperation levels. Related to Bland and Nikiforakis (2015), Chmura *et al.* (2005) show that inequality influences equilibrium selection in coordination games, but this is “best explained not by a preference for equality per se but rather by the belief that the opponent has such a preference”.

³ As our theoretical analysis reveals, *ex-ante* normative conflict can also lead to feuds. In other words, the feuds between heterogeneous agents documented in Nikiforakis *et al.* (2012) need not reflect a lasting disagreement between them, but rather an attempt to “communicate” after failing to coordinate. Individuals are known to use costly punishment to “communicate” when there are no explicit communication opportunities (e.g., Nikiforakis and Mitchell 2014, Noussair and Tucker 2005, Reuben and van Winden 2008, Xiao and Houser 2005).

Our data support the hypothesis that communication and punishment perform different, valuable functions. When introduced early, communication allows nearly 90% of the groups to establish a *covenant*, i.e., a mutual understanding among group members about contributions.⁴ Even though we find evidence that group members often hold different normative views, i.e., that *ex-post* normative conflict is present, the prevalence of covenants indicates that most groups are able to resolve this type of conflict in our experiment. Punishment, on the other hand, ensures that the covenants are virtually always followed. Sanctioning in groups with covenants is about 95% lower than in groups without one, indicating the high costs associated with *ex-ante* normative conflict that can be avoided through communication. The result is that cooperation and earnings are significantly greater when both communication and punishment are possible than when either is unavailable. Even then, however, cooperation levels and group earnings are well below the social optimum, as about 80% of the covenants do not prescribe efficient cooperation. Instead, they typically require less-than-maximal contributions from the agents who benefit relatively little from cooperation. Taken together, these results suggest that cooperation levels fall short of the maximum because many individuals are concerned about equality.

The time at which communication becomes available is critical for its efficacy. Although late communication reduces the overall extent of punishment applied from the level prevailing before communication is introduced, sanctioning remains substantial for many groups. The reason is that prior experiences of sanctioning and feuding are associated with a reduced likelihood of groups establishing covenants later on. Communication is therefore relatively ineffective in fully *overcoming* a history of sanctioning, especially if a feud (i.e., a long sequence of reciprocal sanctions) has occurred. Throughout the session, cooperation and earnings are greater when communication is introduced early rather than late, suggesting that covenants should precede ‘swords’ if possible. This finding reveals limits in the ability of peer-to-peer communication to promote cooperation in groups that have experienced a history of conflict. While the more fundamental problem of *ex-post* normative conflict may not be a great obstacle if agents can communicate early and strike up a covenant before sanctions become widespread, it becomes more of an obstacle when there is a prior history of sanctioning.

⁴ Our use of the term covenant is similar to what Hobbes (1960, p.89) calls “covenants of mutual trust,” i.e., when all parties involved make (explicit or implicit) promises about their future actions. Our usage of the term differs from that of Ostrom *et al.*, (1992), who use it to refer to opportunities to communicate under the assumption that individuals use communication to strike up agreements when given an opportunity to do so. This assumption seems plausible in the homogeneous environment of Ostrom *et al.*, given the common interest of agents to cooperate. In our setting with heterogeneous agents, however, communication could well fail to result in a mutual understanding or an explicit agreement.

The reminder of the paper is organized as follows. Section 2 presents the related literature and our experimental design. Section 3 presents a theoretical framework to understand how heterogeneity and normative conflict can affect efficient cooperation. The experimental results are discussed in section 4. Section 5 presents our concluding thoughts.

2. The experiment

2.1 Related literature

Our experimental design builds on earlier studies that have used laboratory experiments to investigate the impact of punishment, communication and heterogeneity in social dilemmas. We offer an overview of the main findings in this literature. We focus our attention on the four studies that are closest to ours. These explore the impact of punishment and/or communication on cooperation in heterogeneous groups.

In pioneering studies, it has been shown that, *in homogeneous groups*, peer punishment *alone* can help sustain cooperation in social dilemmas, often at the highest level possible (Yamagishi, 1986, Ostrom *et al.* 1992, Fehr and Gächter 2000, Sutter *et al.* 2010). Although it usually increases contribution levels substantially, the impact of punishment on group earnings is less clear due to the associated expenditures on sanctions (e.g., Egas and Riedl 2008, Nikiforakis and Normann 2008, Cason and Gangadharan 2015). This holds true even in the long run when retaliation for punishment can occur, as is the case in our experiment (Engelmann and Nikiforakis 2015). Communication, however, unfailingly increases both cooperation and group earnings (Isaac and Walker 1988, Ostrom *et al.* 1992, Brosig *et al.* 2003, Bochet *et al.* 2006, Bochet and Putterman 2009, Janssen *et al.* 2010, Cason and Gangadharan 2016).⁵ Communication between homogeneous agents can be so effective that adding punishment often does not lead to better outcomes (Bochet *et al.* 2006, Janssen *et al.* 2010, Cason and Gangadharan 2016), although this is not always the case (Ostrom *et al.* 1992, Bochet and Putterman 2009, Andrighetto *et al.* 2013). In any event, the joint use of communication and punishment typically allows homogeneous groups to reach fully efficient outcomes.

Cooperation between heterogeneous agents has not been studied nearly as extensively as that between homogeneous agents. Heterogeneity has been introduced into the endowments (e.g., Dickinson and Isaac 1998, Dickinson 2001, Buckley and Croson 2006, Weng and Carlsson 2015),

⁵ The form of communication differs across studies. Early studies allowed for face-to-face communication. More recently, computer chat-room communication at different points in the experiment has become more common. A feature of chat-room communication is that it can maintain subject anonymity. Bochet *et al.* (2006) show that periodic chatroom communication is as effective as face-to-face communication in promoting cooperation when players are homogeneous.

the returns from the public account (e.g., Fisher *et al.* 1995, Reuben and Riedl 2013), and the productivity of individuals' contributions (e.g., Tan 2008, Noussair and Tan 2011). While the overall effect of heterogeneity on cooperation appears to be unclear (Gangadharan *et al.* 2017), there is consistent evidence that the normative conflict arising when agents obtain different returns from the public account, as in our experiment, is associated with lower levels of cooperation and group earnings. Nikiforakis *et al.* (2012), Reuben and Riedl (2013), and Gangadharan *et al.* (2017) all find that subjects are likely to disagree about which contribution rule should be followed.

The studies closest to ours are those by Nikiforakis *et al.* (2012), Reuben and Riedl (2013), Gangadharan *et al.* (2017), and Dekel *et al.* (2017). Nikiforakis *et al.* (2012) use a set up similar to ours to explore the hypotheses that heterogeneity among agents triggers normative conflict and that normative conflict increases the probability of a feud erupting. In line with this, they find that agents that obtain different benefits from cooperation favor different normative rules. Furthermore, peer punishment is approximately three times as likely to start a feud in heterogeneous than in homogeneous groups. Nikiforakis *et al.* (2012) also find that, while the possibility of a feud sustains cooperation at similar levels in both types of groups, the cost of feuding in heterogeneous groups fully offsets the efficiency gains from increased cooperation. Similarly, Reuben and Riedl (2013) compare outcomes in homogeneous and heterogeneous groups, both in the presence and absence of punishment opportunities. The main difference with Nikiforakis *et al.* (2012) is that retaliation for punishment cannot occur. The authors conclude that groups can overcome the collective action problem even in heterogeneous groups. Forty percent of groups achieve efficient cooperation, while the majority of them seem to coordinate on rules that strike a compromise between efficiency and equality concerns. Importantly, both Nikiforakis *et al.* (2012) and Reuben and Riedl (2013) do not allow group members to communicate. Therefore, we cannot rule out the possibility that the observed inefficiencies are driven by *ex-ante* instead of *ex-post* normative conflict, and that combining communication with punishment opportunities could lead to efficient cooperation.

Gangadharan *et al.* (2017) show that communication (coupled with rewards/monetary transfers) is less effective in heterogeneous than it is in homogeneous groups. Although almost all groups agree to follow specific contribution rules with positive contributions, most of them either prioritize equality over efficiency or strike a compromise between the two. One reason this may happen is because participants have no means to sanction individuals who defect from agreements made. Indeed, the authors observe systematic instances in which individuals first agree on a rule and then proceed to violate it. Interestingly, there was no evidence of individuals using monetary transfers to reduce inequality, suggesting that punishment may be a better instrument to reach efficient cooperation than rewards, at least when conflicting normative views coexist.

The evidence above indicates (1) that communication and punishment *together* usually lead to efficient outcomes in homogeneous groups; and (2) that communication and punishment *alone* are less effective at promoting cooperation in heterogeneous than in homogeneous groups. It remains unclear whether heterogeneous groups can cooperate and maximize efficiency when the instruments are used *in tandem*. The only study in which heterogeneous groups are allowed to use both communication and punishment is by Dekel *et al.* (2017). These authors examine a different kind of social dilemma, in which cooperation confers a benefit to the group on aggregate, but harms a minority of group members (*i.e.*, one of the three group members). Dekel *et al.* (2017) explore the efficacy of punishment, both with and without communication. They find that contributions to “potential Pareto public goods” are not viewed as unequivocally socially desirable, and do not increase either with communication or punishment. An important difference from our study is that minority players receive a negative return from the “potential Pareto public good”, changing fundamentally the incentives of the different types of players, their interaction, and the very nature of the normative conflict, as different normatively appealing rules may be at play. For example, a rule of “do no harm unto others” could make majority players unwilling to contribute to the public good, unlike in our setting where such a norm does not exist. We would expect communication and punishment to be more effective in our setting.

2.2 The basic game

Our basic game is a voluntary contribution mechanism, with a linear production technology for the public good, and the feature that agents benefit differentially from cooperation, all else equal. At the beginning of each period, each participant is given an endowment of 20 experimental currency units (ECU) and must divide it between a “private account” and a “public account”. The earnings of group member i at the end of the first stage are given by:

$$\pi_i^1 = 20 - c_i + m_i \sum_{j=1}^n c_j,$$

where c_i denotes contributions to the public account, $c_i \in \{0, 1, \dots, 20\}$, and m_i is the return to total group allocations towards the public account. This return differs among subjects, $m_i = \{0.3, 0.6\}$. In our experiment, we create groups of four ($n = 4$), with two high-return (and two low-return) players receiving a benefit of 0.6 (0.3) from the public account. (In the instructions, high-return and low-return players are called “type A” and “type B”, respectively.) Group composition and roles remain fixed throughout the experiment. Since $m_i < 1$, while $n \cdot m_i > 1$, the situation poses a social dilemma.

An interesting aspect of heterogeneous groups is that, even if group members generally agree to contribute to the public account, they may still disagree about how much each type should contribute. Since $n \cdot m_i > 1$, the contribution plan that maximizes efficiency, the *efficiency rule*, is $c_i = 20$, for all i . As we will outline in more detail below, in our experiment, efficiency can also follow from a motivation to avoid *inequity*, i.e. a desire to respect that different benefits from cooperation are earned. The *efficiency rule* maximizes overall (and high-return) earnings, but also leads to a high level of inequality between different players ($\pi_{high}^1 = 48, \pi_{low}^1 = 24$). For this reason, low-return individuals may favor the *equality rule* that yields the greatest payoff to all parties, where $c_{high} = 20, c_{low} = 5$ ($\pi_{high}^1 = 30, \pi_{low}^1 = 30$) (Nikiforakis *et al.* 2012, Reuben and Riedl 2013).⁶ Of course, subjects may also strike a *compromise* between equality and equity (implying efficiency), e.g., $c_{high} = 20, 5 < c_{low} < 20$. In principle, groups could also agree on rules that require high-return players to contribute less than 20, but such agreements are rare.

To increase the relative appeal of efficiency as a normative rule (i.e., to make it an equitable outcome), as in Nikiforakis *et al.* (2012) and Gangadharan *et al.* (2017), the different return rates of the public account are assigned based on participants' performance in a real-effort task.⁷ The non-random assignment of return rates is expected to make inequality more acceptable (e.g., Cappelen *et al.* 2007, Gee *et al.* 2017). The task, taken from Erkal *et al.* (2011), is performed once at the beginning of the experiment. Subjects are given a table assigning numbers to letters of the alphabet and are asked to use it to encrypt a number of words for ten minutes. Participants know only that the two best performing subjects in their group will receive a higher return from a "public account". Thus, they understand that good performance in the task will translate into an advantageous position in the subsequent interaction, but nothing else. This lack of information is intended to limit selection effects (Erkal *et al.* 2011). No other monetary incentives are provided for the task. After the encryption task, subjects learn whether they are a high-return or a low-return player ("Type A" or "Type B"), but are not informed about the exact number of words the other members of their group encrypted until the end of the experiment.

⁶ There are a number of contribution profiles that lead to equal earnings to all parties. Each profile in which high-return players contribute four times the amount that low-return players do leads to equal payoffs. However, given that players' goal is to achieve equality, there is no reason for them not to choose the profile that generates the highest level of earnings, which is $c_{high} = 20, c_{low} = 5$. The OSM presents evidence that at least when communication opportunities are present, experimental subjects seem not to have a problem to coordinate on the Pareto-superior equilibrium.

⁷ Notably, the real-effort task does not yield any direct pecuniary benefits to participants or the experimenter, and could be considered as unproductive. Due to its structure, participants coding a lot of words actually exert a negative externality on those coding fewer words, who have to exert greater effort to keep up. We believe that the task induces feelings of entitlement and provides a justification for why some players demand higher earnings than others. In the field, an assessment used for hiring people also does not produce direct benefits. Nonetheless, it provides a justification why some people earn more (i.e., get the job) than others.

2.3 Treatments

There are three treatments in the experiment, each of which consists of two parts. After the real effort task, subjects play 10 periods of the previously outlined (and appropriately modified) public-good game both in *part 1* and in *part 2* of the experiment. In each session only one treatment is in effect. Thus, a session consists of the encryption task, followed by 20 periods of play of the public-good game, divided into two 10-period segments, called *parts* of the session. When in *part 1*, subjects know that a second part will follow, but they do not know what the upcoming part will entail.

In the first two treatments, *ComEarly* and *ComLate*, the setting described in section 2.2 is extended by providing participants with punishment opportunities, as outlined in more detail below. In addition, in these two treatments, we allow subjects to communicate in one part of the session, and we vary the timing of communication between the two treatments. In *ComEarly*, we introduce communication early, in part 1. In *ComLate*, we introduce communication only in part 2. We keep the number of communication opportunities constant over the totality of the session, by disabling communication in part 2 of *ComEarly*. Finally, we implement a treatment with only communication and no punishment opportunity, *ComOnly*. In this treatment, communication is available in both parts 1 and 2. The three treatments are described in more detail below.

2.3.1 *ComEarly*

Subjects are provided with punishment opportunities in both parts. Each period includes the contribution stage of the basic public-good game but also *one or more* additional punishment stages. The number of punishment stages is endogenously determined. In each of these stages, individuals simultaneously decide on how much to punish each other. After they observe the individual contributions and resulting provisional earnings of each group member, subjects have the opportunity to assign punishment points that reduce other players' earnings. If any player assigns punishment in a given stage, another punishment stage follows. If no one assigns points, the period ends and a new one commences.

To assign punishment points, participants have to pay a flat fee of 1 ECU. Payment of the fee allows an individual to assign as many points to as many group members as she wishes for the remainder of the current period.⁸ Each punishment point reduces the earnings of its recipient by 1

⁸ This cost structure has the consequence that low-return and high-return players have a similar ability to punish. Although nominally this punishment technology is cheap, the actual cost is endogenously determined and depends on the desire of punishment victims (who also have access to the same technology) to counter-punish. The available evidence suggests that the severity of counter-punishment increases with that of punishment (e.g., Denant-Boemont et al. 2007, Nikiforakis 2008).

ECU. Let p_{ij}^s denote the number of punishment points that player i assigns to j in punishment stage s (where $i, j = 1, \dots, 4; i \neq j$). Player i 's earnings at the end of punishment stage s are, accordingly,

$$\pi_i^s = 20 - c_i + m_i \sum_{j=1}^n c_j - \sum_{s=1}^S \sum_{j=1}^n p_{ji}^s - \text{Punishment Fee}$$

We refer to the total punishment applied over all punishment stages by an individual i in a given period, $\sum_{s=1}^S \sum_{j=1}^n p_{ij}^s$, as the amount of *sanctioning* he applies. This sanctioning is limited, insofar as the maximum number of points that player j can assign to player i in the first punishment stage is π_i^1 . The punishment points assigned in any stage s must satisfy $p_{ji}^s \leq \pi_i^s$. Thus, earnings can be reduced at most by others to 0. Subjects could always assign points even if the fee paid to sanction others would make their own earnings negative. Permitting such punishment implies that subjects cannot preempt retaliation by handing out severe punishment. If player i receives points from others equal to her entire earnings from the contribution stage and pays the punishment fee, her earnings are -1 ECU. These are the lowest possible period earnings.

A period ends and a new one begins if (a) either no points are distributed in a given punishment stage, or (b) all players' period earnings have already been reduced to 0. At the start of every new punishment stage, participants are informed about the number of points each group member has assigned to them in the previous stage. They also observe the total number of points each group member has assigned to them so far in the period, and the number of points each group member assigned to her peers in the previous stage. The presentation format ensures that players are able to connect their group members' contributions and their subsequent punishment behavior.

Communication is available at the outset of play in *ComEarly*. Our operationalization of communication is inspired by Bochet *et al.* (2006). Anonymous free-form, chatbox communication is implemented three times during the first 10 periods: at the start of periods 1, 4, and 7. Notably, this also implies that subjects can communicate with the other group members before any punishment has been applied.⁹ Allowing communication in only every third period saves time and has been found to be as effective as more frequent communication in creating cooperation in a homogenous environment (Bochet *et al.* 2006). The chat rooms are open for 5 minutes at the start of period 1 and for 3 minutes in the other communication stages. We disable the communication in

⁹ Subjects can actually use two different chat boxes. One box allows the sending of messages to all other group members while the other only allows messages to be sent to the group member of the same type. Communicating with the whole group can potentially foster cooperation by helping to reach a compromise. At the same time, the opportunity to chat with the same-type group member may lead low (or high) types to sabotage a group-wide compromise. See OSM for a detailed discussion on how participants use the boxes to communicate.

part 2 of *ComEarly* in order to have an equal number of communication opportunities in the two main treatments, and to study the persistence of any effects of communication after it is no longer possible.

2.3.2 *ComLate*

As in *ComEarly*, subjects have punishment opportunities in both parts 1 and 2 of the experiment. In *ComLate*, however, communication is introduced relatively *late*, and is only available in part 2 of the session (periods 11, 14, and 17). Given the results of Nikiforakis *et al.* (2012), we anticipate that normative conflict will lead to feuds and high levels of overall sanctioning in the first ten periods. This treatment therefore allows us to study the extent to which communication is able to overcome a history of sanctioning that materialized before any communication was possible.

2.3.3 *ComOnly*

The previous treatments switch communication on and off, allowing us to compare outcomes when communication and punishment are both present *vs.* when only punishment is present, either in the first or the second part. The *ComOnly* treatment switches punishment off in both parts of the session, providing us with a benchmark for evaluating the impact of punishment in the previous treatments. In order to be able to compare outcomes when both communication and punishment are present *vs.* when only communication is present *in both parts* (i.e., compare *ComEarly* *vs.* *ComOnly* in part 1, and *ComLate* *vs.* *ComOnly* in part 2), subjects can communicate at the start of periods 1, 4, and 7, in part 1, and periods 11, 14, and 17, in part 2.

2.4. *Procedures*

Table 1 provides a summary of the experimental treatments and the number of independent groups in each. In order to collect a sufficient number of independent observations, the sessions were conducted in two locations: the CentERlab at Tilburg University, and at the SSEL at NYU Abu Dhabi.¹⁰ Between 12 and 24 subjects participated in each session. Recruitment was done via h-root (Bock *et al.* 2014) at NYU Abu Dhabi and with an internally programmed software program at Tilburg. None of the 204 subjects participated in more than one session.¹¹ The experiment was

¹⁰ The subject pool at NYU Abu Dhabi is ethnically mixed and not unlike that in other universities in many developed countries. Since there is no *a priori* reason to expect any differences between the two subject pools, we pool the data in our analysis, and control for location in our regressions. Nevertheless, we note that we find no significant differences in contributions, earnings, or sanctioning levels between the two locations, using non-parametric tests.

¹¹ We exclude one group from *ComEarly* from our analysis. The communication analysis revealed that one group member thought it would be necessary to punish by one point in order to reach the next punishment stage, which she

run using z-Tree (Fischbacher, 2007). At the beginning of part 1 of the experiment, the instructions were distributed and read aloud by the experimenter to ensure that they were common knowledge. Subjects had to answer an extensive control questionnaire before the experiment could start. Updates to the instructions, indicating any changes that were about to come into effect, were distributed and read aloud before part 2.

Table 1 – Number of groups, by treatment and location

Treatment	Part 1	Part 2	Number of groups		
			Abu Dhabi	Tilburg	Total
<i>ComLate</i>	Punishment	Punishment & Comm.	7	11	18
<i>ComEarly</i>	Punishment & Comm.	Punishment	6	10	16
<i>ComOnly</i>	Comm.	Comm.	6	11	17

The sessions lasted approximately 105 minutes in *ComOnly* and 150 minutes in *ComEarly* and *ComLate*. Show up fees and conversion rates in both locations were adjusted to reflect the earnings typically paid at the two laboratories. In Tilburg, subjects earned on average 32.00 EUR (= \$36US), including a show-up fee of 5 EUR, where 1 ECU was valued at to 0.045 EUR. In Abu Dhabi subjects earned 280 AED on average (approximately 62 EUR = \$69US), including a show-up fee of 30 AED, and the conversion rate was 1 ECU to 0.45 AED. In terms of ECU, experimental earnings were similar in the two locations (5.9% higher in Tilburg).

3. Theoretical framework and hypotheses

To derive hypotheses for our experiment, we use a simple theoretical framework to help delineate the conditions under which heterogeneous benefits from cooperation can lead to normative conflict and, more importantly, when normative conflict might be an obstacle to efficient cooperation. In order for our framework to provide useful insights, it must be able to account for the findings in previous studies with heterogeneous groups. Specifically, the model must be able to account for the incidence of feuds seen in Nikiforakis *et al.* (2012), and the tendency of individuals to undercut contribution agreements in the absence of a punishment threat in Gangadharan *et al.* (2017).

thought to be beneficial. After the other group members explained her mistake to her in a communication stage, she stopped punishing. Including this group affects none of our main results, and only increases the number of instances classified as feuds in this treatment.

Our focus is on the normative conflict that arises when there is a tension between equality and efficiency. To that end, we model normative conflict by considering three distinct behavioral types: selfish money maximizers, inequality-averse and inequity-averse individuals.¹² Selfish money maximizers are assumed to have preferences given by $u_i(\pi) = \pi_i$. We follow Fehr and Schmidt (1999) in modeling the preferences of those who dislike *inequality*. The utility of such an individual i is given by:

$$u_i(\pi) = \pi_i - \alpha_i \frac{1}{n-1} \sum_{j \neq i} \max(\pi_j - \pi_i, 0) - \beta_i \frac{1}{n-1} \sum_{j \neq i} \max(\pi_i - \pi_j, 0) \quad (1)$$

where α_i measures the disutility i derives from disadvantageous inequality and β_i the disutility s/he derives from advantageous inequality, with $\alpha_i \geq \beta_i > 0$, and $\beta_i \leq 1$. Naturally, inequality-averse player will find the *equality rule* attractive. We assume that *inequity*-averse individuals have a similar utility function as inequality-averse individuals with the difference that their reference payoff is weighed to reflect equity considerations. In particular, we assume that the utility of such an individual is given by:

$$u_i(\pi) = \pi_i - \gamma_i \frac{1}{n-1} \frac{1}{m_j} \sum_{j \neq i} \max(m_i \pi_j - m_j \pi_i, 0) - \delta_i \frac{1}{n-1} \frac{1}{m_i} \sum_{j \neq i} \max(m_j \pi_i - m_i \pi_j, 0) \quad (2)$$

where $\gamma_i \geq \delta_i > 0$, and that $\delta_i \leq 1$. Equation (2) implies that someone who has, e.g., twice as high a return from the public good should earn twice as much. If $m_i > m_j$ then i will suffer a loss in utility if her monetary payoff is not sufficiently larger than that of j . The underlying idea of the utility specification is that only when $m_i \pi_j - m_j \pi_i = 0$ there is there no utility loss. In other words, inequity is zero if $m_i/m_j = \pi_i/\pi_j$. Inequity-averse players find the *efficiency rule* attractive: If one accepts that payoffs should be in line with the returns from the public good, the efficient outcome is desirable as high-return players earn exactly twice as much as low-return players when everyone contributes fully (i.e. there is no inequity).

What can we expect to happen when these different behavioral types coexist in a group? Can communication and punishment support efficient cooperation? Below, we provide answers to these

¹² We have opted to use inequity-averse players as opposed to efficiency-concerned individuals in our framework for two reasons. First, as the efficient outcome is an equitable outcome in our experiment, we can simply not distinguish whether players favoring the ‘efficiency rule’ care about efficiency or whether they favor equity. Second, as we discuss in more detail in the OSM, assuming the existence of efficiency-minded as opposed to inequity-averse players would not allow for feuds as in Nikiforakis *et al.* (2012) as an efficiency-minded person never reduces another’s earnings. We discuss other modeling approaches involving concerns for efficiency in the OSM, but inequity aversion seems a natural way to capture the idea that high-return players may insist on efficient contributions from everyone.

questions and hypotheses for our experimental treatments using the insights obtained from our theoretical analysis; formal propositions and details of the model can be found in the OSM.

The potential coexistence of different behavioral types in any given group implies that individuals are likely to be uncertain about the preferences of others in their group, a phenomenon we call *ex-ante normative conflict* (EANC). *Communication* can serve to reduce this uncertainty and, in our framework, for simplicity, we assume that it fully removes the uncertainty (transforming a game of incomplete information to one of complete information).¹³ Relatedly, we define *ex-post normative conflict* (EPNC) as a situation in which it is common knowledge that inequity and inequality averse types coexist in a group.^{14,15} *Punishment* provides a coercive force that inequality-averse and inequity-averse players can use to find a compromise among themselves (and motivate money-maximizing players to contribute).

What should we expect to happen in our three experimental environments (Communication alone, Communication with Punishment, and Punishment alone)? When communication is possible but punishment is not, we establish (along the lines of previous results in the literature) that the equality (efficiency) rule can be maintained in equilibrium only if *all* group members are sufficiently inequality (inequity) averse, i.e., there is no normative conflict. Otherwise – if at least one member favors equality while at least one other member favors equity – cooperation unravels. Similarly, one selfish player is sufficient to lead cooperation to unravel. Thus, communication alone is unlikely to lead to efficient cooperation.

When both communication and punishment are possible, players can punish deviations from their preferred contribution rules. This changes the situation fundamentally. Without conflicting normative views, the equality (efficiency) rule can already be maintained by one sufficiently inequality- (inequity-) averse player who punishes free-riding selfish players. With EPNC, we establish that there are many situations of *resolvable EPNC* – instances when EPNC does not lead

¹³ This is clearly a strong assumption. There is, of course, ample evidence showing that many individuals incur psychological costs when lying (Lundquist *et al.* 2009, Gneezy *et al.* 2018). Furthermore, in our setting with fixed IDs, individuals have an incentive to maintain a reputation of being perceived as honest. We recognize, however, that some players might still be willing to lie, especially in the absence of punishment opportunities. Nevertheless, we abstract from this scenario in our analysis for tractability, since the possibility of some players lying only reduces the predicted efficacy of communication but does not affect the comparative statics in our analysis.

¹⁴ In line with the evidence in Nikiforakis *et al.* (2012) and Reuben and Riedl (2013), we assume that high-return players in such a situation exhibit concerns for inequity, while low-return players for inequality, i.e., individuals select the normative rule to favor in a self-serving manner.

¹⁵ Our definitions imply that the coexistence of either inequality or inequity averse players with the selfish type does not qualify as normative conflict. The underlying reason is that while in such a situation, the normative rules of the two player types might be different, the selfish players will always adjust their behavior if threatened by punishment. Selfish players do not view selfish behavior as a norm that the group should adhere to.

to the unraveling of cooperation – and some of *unresolvable EPNC* – instances when EPNC leads to an unraveling of cooperation and/or possibly sanctions and feuds. Intuitively, when the dislike of both inequality and inequity are not too strong among players, compromises are possible.¹⁶ Importantly, the efficient outcome is attainable as long as the aversion against inequality is not too high. In contrast, when the distaste for both inequality and inequity among players is high, normative conflict cannot be resolved, and will lead to an unraveling of cooperation and/or sanctioning. In extreme cases, this can imply a full destruction of payoffs due to sanctions (e.g., from feuding). Nonetheless, as the requirements for latter cases seem fairly high, we would expect instances of feuding to be rare when punishment and communication opportunities coexist. Therefore, we would expect groups to be better off overall than under communication alone.

Finally, when communication is not possible, opportunities to punish others can still enable groups to compromise even when there are conflicting normative views. However, the presence of uncertainty – or ex-ante normative conflict (EANC) – makes cooperation more difficult, in particular when conflicting normative views exist (EPNC). We establish that with uncertainty, low-return players may sometimes contribute insufficiently due to their beliefs, which in turn can lead to additional sanctioning and even very costly feuds in some cases as seen in Nikiforakis *et al.* (2012). Overall, these insights lead to our first testable hypothesis:

Hypothesis 1. [*Communication and Punishment*]: *Punishment and communication together will lead to higher contributions and earnings when jointly present, than when either communication alone or punishment alone is in effect. In addition, punishment will be greater when no communication is possible than when it is.*

The fact that efficient cooperation is attainable when communication and punishment are available implies that ex-post normative conflict *can be* resolved in favor of efficiency, as long as individuals are not too strongly averse to inequality. Failure to observe efficient cooperation in this case, therefore, could either be due to *unresolvable* ex-post normative conflict or to groups not regarding equity/efficiency as a sufficiently attractive normative rule. To evaluate the importance of each factor, we empirically explore whether groups manage to strike “covenants” and their content. Failure to reach covenants would be clear evidence of unresolvable ex-post normative conflict.

¹⁶ Inequity-averse players prefer the efficiency rule but might be willing to accept lower contributions from low-return players, $c_{low} < 20$, as long as their aversion against (disadvantageous) inequity is not too high. Similarly, inequality-averse players prefer the equality rule but might be willing – under the threat of punishment – to contribute $c_{low} > 5$, as long as their aversion against (disadvantageous) inequality is not too high.

How should we expect the timing of communication to impact the ability of groups to cooperate efficiently? We do not offer a formal theoretical exploration for this question, however, we can provide some intuitive reasoning to guide our analysis. Communication should greatly enhance participants' understanding about the normative views of their group members. Removing the opportunity to communicate in the second part of *ComEarly* should, therefore, have little impact on cooperation, and any agreements agreed to earlier should continue to be followed. The introduction of communication in *ComLate* should alleviate EANC, presumably to the same extent it did in the first part of *ComEarly*, but only beginning in the second part of the session. Our hypothesis regarding the timing of communication is the following:

Hypothesis 2. *[Timing of communication]: Over both parts of the sessions, contributions and earnings are higher in ComEarly compared to ComLate. The difference is driven by the first ten periods.*

4. Results

The discussion of our findings is separated into three subsections. The first explores whether communication and punishment can lead to efficient cooperation when acting in tandem, and the distinct roles of EANC and EPNC. The second subsection explores whether the timing of communication affects its efficacy. Finally, the last subsection analyses the content of the covenants that are established.

4.1. Communication and punishment: together and alone

To test Hypothesis 1, we use the data from the first part of the experiment, which allows for a comparison that is unaffected by differences in past outcomes.

Result 1 [Contributions & earnings]: *Contributions and earnings are highest when communication and punishment are available in tandem. Hypothesis 1 is supported. Even then, however, contributions and earnings are well below the social optimum.*

SUPPORT: Figure 1 shows the evolution of average contributions and earnings in each treatment in the first part of the experiment. Average contributions are substantially higher in *ComEarly* (15.19) – when both communication and punishment are possible – than in *ComLate* (11.94) and *ComOnly* (11.26), where only one of the instruments is available (Mann-Whitney, two-tailed, $p < 0.032$ for both comparisons). Contributions are roughly 25% below the maximum in *ComEarly*. On average, earnings are also significantly higher in *ComEarly* (31.45) than in *ComLate* (24.34) and *ComOnly* (29.01) (Mann-Whitney, two-tailed, $p < 0.040$ for both comparisons). Earnings in *ComEarly* are roughly 87% of the maximum. While contributions are not significantly different

between *ComLate* and *ComOnly* (Mann-Whitney, two-tailed, $p > 0.20$), earnings are marginally higher in *ComOnly* (Mann-Whitney, two-tailed, $p = 0.098$).¹⁷ ■

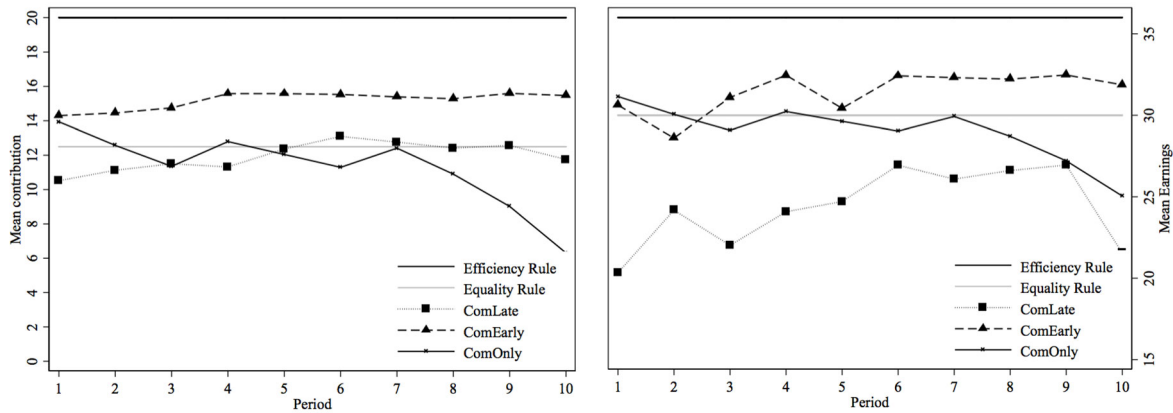


Figure 1 – Average Contributions (left) and Earnings (right) in Part 1, by Treatment

To better understand the effect of communication on cooperation, we turn our attention to *covenants* which, as mentioned previously, refer to a mutual understanding, resulting from communication among group members, about how much each person will contribute to the public account. A covenant can be either *explicit* or *implicit*. We say that an *explicit covenant* has been established if all group members actively state their agreement to follow a specific contribution rule. We say that an *implicit covenant* has been created if there is no active disagreement between group members (i.e., there is a *mutual understanding*) about which contribution rule would be followed.¹⁸ For brevity, unless otherwise specified, we will subsume both *explicit* and *implicit* covenants under the term *covenant*.¹⁹ We will also say that a group *adheres* to a covenant if all of

¹⁷ Table A1 in the OSM also displays the mean contributions and earnings by treatment and part of the session, as well as non-parametric tests of the treatment differences.

¹⁸ To determine whether a group has established a covenant, we use the transcripts of the group communication. The data was classified by three independent coders (see OSM for details, including an analysis of coder reliability). In cases where no explicit covenant was established, we asked our three coders to rank the level of disagreement in each group on a scale from 0 to 4. A score of 0 is indicative of a mutual understanding, as we asked coders to rate the discussion in a group “in which the fourth person just fails to agree to a contribution rule because time has run out” with a 0. In contrast, we suggested rating “a group that discusses alternative rules and cannot agree on one rule but in which players remain polite during the whole discussion” as a 2. In addition, we required the level of disagreement in a group to be rated as 3 or 4 if individuals threaten to punish others, insult one another, or behave in similarly hostile ways. Following a conservative approach, we say that a group has established an implicit covenant if at least two coders assigned the lowest possible level of disagreement to the group, 0, and no coder assigned a level greater than 1. The determination of whether a covenant exists in a given group was done separately for each of the individual communication periods, since the content of covenants can vary across communication periods.

¹⁹ Most covenants – about 80% – are *explicit*. We arrive at similar results when using only explicit covenants in our analysis.

its members *strictly* follow a contribution rule in the next three periods after the communication stage in which a rule was agreed upon (i.e., until the next communication stage).

Result 2 [Covenants & adherence]: *Communication enables most groups to establish covenants. The threat of punishment increases the likelihood that groups adhere to their covenants.*

SUPPORT: In Part 1 of both *ComEarly* and *ComOnly*, there were three communication stages. In *ComEarly*, in 87% of possible instances, groups establish a covenant. The same is true for 81% of cases in *ComOnly* (χ^2 test, two-tailed, $p > 0.20$).²⁰ While groups adhere to their covenant in 81% of the cases in *ComEarly*, they do so in only 56% of instances in *ComOnly*, where the threat of punishment is not present (χ^2 , two-tailed, $p = 0.010$). ■

Result 3 [Covenants & sanctioning]: *Covenants reduce sanctioning by 95%.*

SUPPORT: Figure 2 shows the average number of punishment points an individual assigns in a period, totaled across all punishment stages, in part 1 of *ComEarly* and *ComLate*. In *ComEarly*, we distinguish between sanctioning in groups that have established a covenant and those who have not. Sanctioning expenditure in *ComEarly* is on average 95% lower in groups with covenants than it is in groups without covenants (0.22 vs. 4.13 points per individual in a period; $p = 0.001$).²¹ Punishment is similar in groups without a covenant in *ComEarly* (4.13 points) and groups in *ComLate* (5.42 points), where covenants are prevented by design ($p = 0.701$). Since most groups establish a covenant in *ComEarly*, the overall level of punishment in *ComEarly* (0.68 points) is substantially and significantly lower than in *ComLate* ($p = 0.004$). ■

²⁰ To ensure that observations are independent for the χ^2 test, we use only observations from one communication period for each group. To abstract from initial coordination problems and limit potential end-game effects, we use only data from the second communication period.

²¹ Throughout the paper we utilize random-effects regressions to analyze the significance of differences in sanctioning, unless otherwise stated. This allows us to control for group contributions. The dependent variable in these regressions is the number of points player i assigns to player j over all punishment stages. Simple non-parametric tests generally yield very similar results.

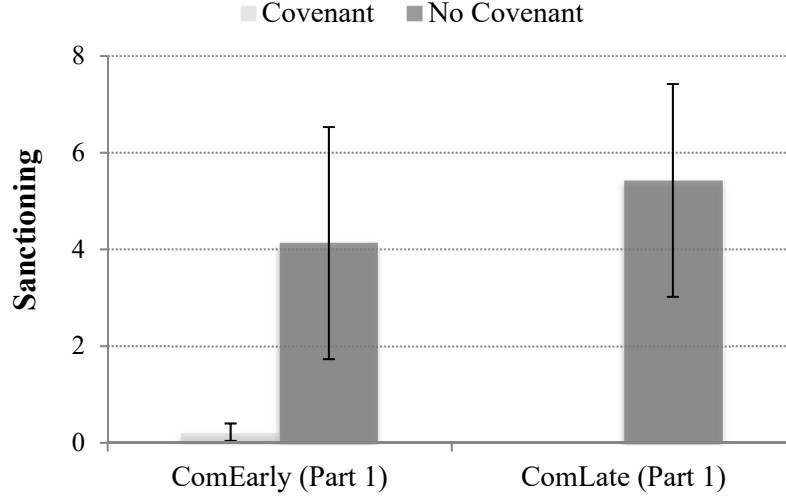


Figure 2 – Average Quantity of Sanctions in part 1: Covenants vs. No Covenants in *ComEarly*; No Covenants in *ComLate* (95%-Confidence Intervals)

Results 2 and 3 provide some first evidence that unresolved EPNC (as observed in *ComEarly*) is a less substantial problem than EANC (as observed in *ComLate*) is in our experiment.²² One may be worried about reverse causality in Result 3. Rather than covenants reducing punishment expenditure, it may be that covenants are more likely to be established in groups with low levels of sanctioning. Indeed, we will see later in Result 5 that a history of heavy sanctioning can undermine the ability of groups to reach covenants. However, this does not affect our claim in Result 3 that covenants reduce sanctioning. If we restrict our analysis to the first three periods, i.e., after the first communication instance which occurs prior to any sanctioning, we obtain very similar results: 80% of the groups in *ComEarly* establish covenants. The average punishment assigned in these groups is 0.60 points per period, compared to 4.80 points in groups without covenants ($p = 0.006$).

4.2. *The timing of communication: covenants before the swords*

To investigate how critical the timing of communication is, we use data from *both parts* of the experiment. This offers the cleanest comparison as it controls for the total number of

²² In line with these observations, the main reason for the differences in earnings between the two treatments is the use of punishment when no covenant has been made, which accounts for 64.2% of the difference. We can disaggregate the extent to which the difference in earnings between *ComEarly* and *ComLate* is due to differences in sanctioning vs. differences in contributions. In the aggregate, earnings in *ComLate* are 22.6% lower than in *ComEarly* (31.87 vs. 24.34). If no punishment had been meted out in either of the treatments, mean earnings would be only 8.1% lower in *ComLate* (32.15 in *ComEarly*, and 29.56 in *ComLate*)

communication rounds across treatments (three in the first part of *ComEarly* and three in the second part of *ComLate*). We also discuss the data from the second part in isolation.

Result 4 [Timing & sanctioning]: *The level of sanctioning in ComLate is more than eight times greater than in ComEarly, over the two parts of the session. Despite the introduction of communication in ComLate, sanctioning expenditure in the second part of the sessions is nearly ten times as high as in ComEarly.*

SUPPORT: Figure 3 shows the amount of sanctioning, over both parts of the session, in *ComEarly* and *ComLate*. On average, an individual assigns only 0.51 points per period in *ComEarly*, compared to 4.31 points per period in *ComLate*; a highly significant difference ($p = 0.005$).²³ The difference between treatments persists in the second part of the session, after communication has been introduced in *ComLate* and disabled in *ComEarly*, where it remains the case that there is more sanctioning in *ComLate* (3.20 points) than in *ComEarly* (0.34 points, $p = 0.063$). The difference in sanctioning between part 2 in *ComLate* (3.20 points) and part 1 in *ComEarly* (0.68 points) is also (weakly) significant ($p = 0.100$), though this last comparison does not control for possible learning effects. ■

To understand why late communication is limited in its capacity to overcome a history of sanctioning, we turn our attention to *feuds* and how they influence the likelihood of establishing covenants. Feuds are sequences of reciprocal sanctioning actions and as such they are of interest as they reveal unresolved tensions within groups. Nikiforakis and Engelmann (2011) say a feud has occurred in period t , if there exists a punishment stage s in period t , for which $p_{ij}^s > 0$, $p_{ji}^{s+1} > 0$, and $p_{ij}^{s+2} > 0$ when feasible, where p_{ij}^s is the number of punishment points that player i assigns to player j in stage $s > 0$. For $s > 1$, we require the additional condition that $p_{ji}^{s-1} = 0$. For $s > 1$, we require the additional condition that $p_{ji}^{s-1} = 0$. That is, a feud is defined as an episode in which punishment is applied, there is retaliation in the next stage, and the original punishing party sanctions the counter-punisher again in the subsequent stage.²⁴

²³ In *ComLate*, over both parts, roughly 60% of sanctioning is meted out in the first punishment stage, while 40% is applied in later punishment stages. In *ComEarly*, the breakdown is 50% each. Most of the higher-order punishment (76% over both treatments) is counter-punishment, i.e., a direct response to prior punishment.

²⁴ In addition, we also consider as feuds the cases in which either (a) stage 1 punishment or (b) stage 1 punishment and stage 2 counter-punishment, are already so severe that a third stage does not even occur, because initial earnings are already fully erased, making a feud in the above-defined sense impossible. This type of feud is also more in line with our modeling approach in the OSM. Notably, we restrict our analysis to within-period feuds, since feuds rarely spill over into the subsequent period.

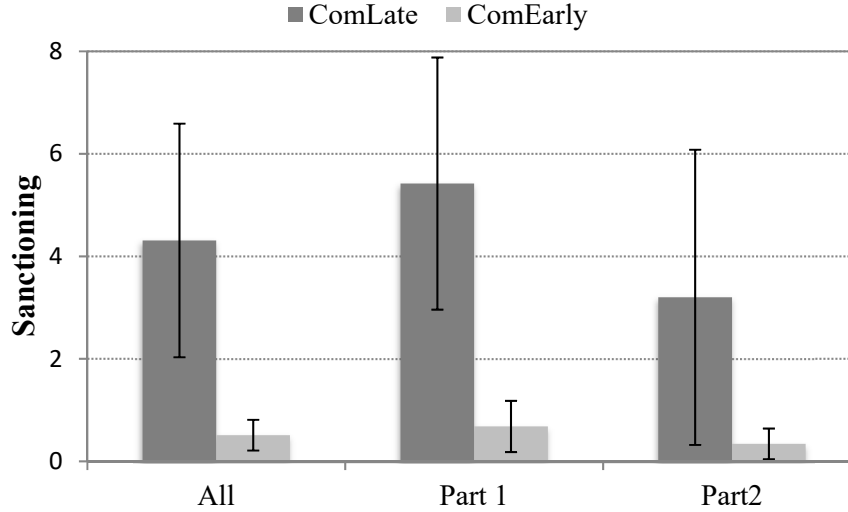


Figure 3 – Average Quantity of Sanctions Assigned in a Period (95%-Confidence Interval)

For feuds, similar patterns emerge as those for sanctioning shown in Figure 3. Overall, 40 feuds occur in *ComLate*, 29 of which take place in the first part of the session, and 11 in the second. Only 6 feuds appear in *ComEarly* (all in the first part of the session). Only 3 out of 16 groups experience at least one feud at any time in *ComEarly*, compared to 10 of the 18 groups in *ComLate* (χ^2 , two-tailed, $p = 0.028$). Similarly, no group experiences a feud in the second part of *ComEarly* compared to 3 groups in *ComLate* (χ^2 , two-tailed, $p = 0.087$). The number of feuds observed in *ComLate* and *ComEarly* suggests that EANC poses a greater problem in our experiment than unresolved EPNC. These data hint that a history of sanctioning can help explain the persistence of sanctioning after communication has been introduced in *ComLate*.

Result 5 [History of sanctioning and covenants]: *Feuding in the first part of ComLate is associated with a reduced likelihood of establishing a covenant in the second part.*

SUPPORT: In *ComLate*, 9 out of 18 groups experienced a feud at least once in the first part of the experiment. For these groups, the likelihood of establishing a *covenant* that lasts through the second part of the experiment is lower than for groups without a history of feuding (55% vs. 88%: χ^2 , two-tailed, $p = 0.114$). This effect is statistically significant when looking at explicit covenants (44% vs. 88%: χ^2 , two-tailed, $p = 0.045$). Overall there are 54 incidences of group communication in the second part of *ComLate* (3 communication stages * 18 groups). Out of the 15 incidences in which no explicit covenant is established, 12 occur in cases when there has been a prior feud in the first part. Moreover, there is a strong positive correlation between the number of punishment points assigned in part 1 and part 2 of *ComLate* (Spearman's $\rho=0.85$, $p = 0.001$). Punishment in part 2 of *ComLate* is almost entirely driven by groups that have experienced a feud in part 1. Members of

groups that experienced a feud in part 1 assigned on average 6.21 points per period in part 2, compared to only 0.29 points for those in groups that did not experience a feud.²⁵ ■

Taken together, these findings indicate that the absence of early communication opportunities is associated with heavy sanctioning and feuds, which are due to EANC. This in turn reduces considerably the likelihood of *some* groups of establishing a covenant, and thereby resolving EPNC, later on, when communication is possible. In the online supplementary material, we provide robustness checks of this result from a regression analysis (see Table A2).²⁶

Result 6 [Timing, contributions & earnings]: *Although the introduction of communication increases both contributions and earnings in the second part of ComLate, overall earnings over both parts of the session are significantly lower in ComLate than in ComEarly. Overall contributions do not differ between the two treatments. Thus, Hypothesis 2 is supported for earnings but not for contributions.*

SUPPORT: The introduction of communication in *ComLate* increases both contributions (from 11.94 to 15.40) and earnings (from 24.34 to 29.72). Both changes are highly significant (Mann-Whitney, two-tailed, $p = 0.001$). Over both parts, earnings (31.66 vs. 27.03) are significantly higher in *ComEarly* than in *ComLate* (Mann-Whitney, two-tailed, $p = 0.014$), even though there is no substantial difference in contributions (15.21 vs. 13.67; $p = 0.207$). (See Table A1.) ■

Note that, despite the higher levels of sanctioning in part 2 in *ComLate* relative to *ComEarly* (see Result 4), earnings are not significantly greater in *ComEarly* than in *ComLate* (Mann-Whitney, two-tailed, $p = 0.414$). The reason is the high variability in outcomes across groups, especially in *ComLate*. Those groups that established a covenant over the whole second part of *ComLate* have average earnings of 32.42, which is similar to the average earnings in *ComEarly* (31.87). On the other hand, groups that did not establish a covenant tended to experience intense sanctioning and had much lower earnings (20.29, Mann-Whitney, two-tailed, $p < 0.010$). It is also notable that

²⁵ Splitting groups at the median level of total sanctioning for the first part of the session leads to similar conclusions. Groups that experienced a degree of sanctioning above the median have a 55% chance of establishing an explicit covenant, while those below the median have a 77% chance. Groups that experienced *higher-order* punishment, that is, punishment in the second stage or later, above the median level have a much lower likelihood of establishing an explicit covenant: 44% vs. 88%. Similar results also emerge when considering all covenants rather than explicit covenants.

²⁶ Despite the failure of late communication in some groups to reduce overall sanctioning, a similar percentage of groups are able to establish a covenant in *ComLate* and *ComEarly* (80% vs. 87%). These covenants appear to keep the level of sanctioning down. In part 2 of *ComLate*, sanctioning averages only 0.21 points in groups with covenants, and 15.21 points in groups without covenants, implying that covenants are associated with a 99% reduction in punishment ($p < 0.001$). Interestingly, sanctioning when no covenant exists is about three times higher in part 2 of *ComLate* than in part 1 when there are no communication opportunities, where it averaged 5.42 points ($p = 0.075$).

subjects in *ComOnly* increase their contributions and earnings in part 2 of the experiment (signed-rank test, two-tailed, both $p < 0.05$). Even though communication has been switched off in part 2 of *ComEarly*, contributions and earnings remain, however, higher than in *ComOnly* (Mann-Whitney, two-tailed, both $p < 0.10$).

4.3. *Covenants and behavior*

Despite the establishment of covenants and the adherence to them under the threat of ‘swords’, contributions remain considerably below the level that would maximize group earnings ($c = 20$ for all group members). Across the two parts, average contributions are only 68% (*ComLate*), 76% (*ComEarly*), and 60% (*ComOnly*) of this benchmark. To understand why average contributions are relatively low, in this section we investigate the content of the covenants, i.e., the contribution rules group follow.

Result 7: *In ComEarly, groups are more likely to adhere to a compromise rule than in ComLate. When punishment is possible, the efficiency and the equality rules are equally likely to be followed. Under ComOnly, efficient contribution rules are rarely employed.*

SUPPORT: Table 2 shows the frequency with which groups follow a specific contribution rule in *ComEarly*, *ComLate*, and *ComOnly*, separated for those groups that establish a covenant and those that do not, in the parts of the sessions in which subjects can communicate. When a covenant has been established, subjects are more likely to follow a compromise rule in *ComEarly* (45%) than in *ComLate* (21%). A compromise rule is a contribution profile that lies between those prescribed by the efficiency and equality rules. Under a compromise rule, high-return players contribute more than low-return players but the difference is smaller than that required to equalize earnings, so that high types do earn more than low ones. According to a χ^2 test, this difference is significant ($p = 0.046$).²⁷ The table also shows that in both *ComLate* and *ComEarly*, the two treatments in which punishment is possible, the percentage of groups following the “efficiency” and “equality” (over both parts) is fairly similar, at approximately 20% each. In *ComOnly*, over both parts, only 10% of groups follow the efficiency rule. ■

The prevalence of compromise rules in our experiment is evidence that low-return and high-return players hold different normative views. Our communication data (see OSM for details) also clearly reveals that different normative rules are mentioned during group chats, with both the

²⁷ As before, we base our tests on the second communication stage to ensure observations are independent and to avoid end-game effects. Since established covenants are also followed most of the time in part 2 of *ComEarly* when communication is disabled, the observed difference also translates into a difference between *ComEarly* and *ComLate* over both parts of the session.

equality and efficiency rules featuring prominently. This EPNC is often resolved by finding a compromise and, thus, many groups fall short of efficient outcomes. In the OSM, we further analyze the variance of earnings within groups and find additional support for the notion that concerns for inequalities in earnings play a significant role and lead groups away from efficient outcomes.

Table 2 – Contribution Rules and Covenants

(a) ComEarly

	<i>Part 1</i>			<i>Part 2</i>	<i>Both parts</i>
	No covenant	Covenant	Weighted av.		
	13%	87%			
Efficiency rule	0.17	0.14	0.14	0.19	0.17
Equality rule	0.17	0.21	0.21	0.23	0.22
Compromise rule	0.17	0.45	0.41	0.56	0.49
No consistency	0.50	0.19	0.07	0.02	0.05

(b) ComLate

	<i>Part 1</i>		<i>Part 2</i>		<i>Both parts</i>
	No covenant	Covenant	Weighted av.		
	20 %	80%			
Efficiency rule	0.11	0.00	0.35	0.28	0.20
Equality rule	0.02	0.27	0.30	0.29	0.16
Compromise rule	0.00	0.09	0.21	0.19	0.09
No consistency	0.87	0.63	0.14	0.24	0.56

(c) ComOnly

Rules	<i>Part 1</i>			<i>Part 2</i>			<i>Both parts</i>
	No covenant	Covenant	Weighted av.	No covenant	Covenant	Weighted av.	
	19%	81%		10%	90%		
Efficiency rule	0.00	0.15	0.12	0.00	0.09	0.08	0.10
Equality rule	0.10	0.15	0.14	0.20	0.30	0.29	0.22
Compromise rule	0.00	0.27	0.22	0.20	0.26	0.25	0.24
No consistency	0.90	0.44	0.53	0.60	0.35	0.38	0.45

Notes: The table shows the frequency with which groups follow different contribution rules (efficiency, equality, compromise) for all periods. Following a rule implies that subjects contribute according to a given rule for three consecutive periods (between communication stages). *Efficiency* refers to full contribution of all group members, *equality* refers to the contribution rule resulting in equal earnings for all group members and *compromise* refers to a compromise between the two other rules. *No consistency* is the residual category. Groups within this category do not (consistently) follow a rule.

4. Conclusion

Can individuals that derive different benefits from cooperation manage to cooperate efficiently when they are provided with two instruments known to be highly effective at promoting cooperation in homogeneous groups? In line with our simple theoretical framework, our experimental evidence indicates that peer-to-peer communication and punishment play complementary roles in heterogeneous groups. Communication enables the vast majority of groups to establish covenants, agreements regarding how much each individual should cooperate, while the *threat* of punishment allows groups to enforce the covenants and maintain cooperation until the very end of the interaction. When the threat posed by punishment is absent, many groups deviate from the covenants that they have agreed upon, leading to a decay in cooperation. As a consequence, cooperation and group earnings are greatest when communication and punishment are both available and can act in tandem. Or, to use Hobbes' (1960) terminology, covenants need to be accompanied by 'swords'.

'Swords', however, must also be accompanied by covenants as these reduce sanctioning and the associated costs drastically. In light of past findings concerning the ambiguous role of punishment at promoting efficiency (e.g., Engelmann and Nikiforakis 2015), especially when there is normative conflict (Nikiforakis *et al.* 2012), communication is surprisingly effective at preventing feuds and limiting the adverse effects of sanctioning. In our terminology, communication seems to prevent ex-ante normative conflict, and punishment opportunities provide the coercive force to resolve most ex-post normative conflict. The timing of communication, however, is critical for obtaining this result. Early communication leads to near-zero levels of sanctioning, as most groups establish covenants. Although late communication leads to improvement in outcomes over those in earlier periods without communication, its efficacy at helping groups overcome a history of feuds is limited. Past feuds undermine the ability of some groups to reach covenants later on. In other words, unimpeded ex-ante normative conflict can influence the resolvability of ex-post normative conflict. This implies that covenants should *precede* the 'swords'.

Scholars interested in finding ways to foster cooperation will find both good news and bad news in our data. The good news undoubtedly is that, with the right timing, communication and punishment opportunities together can help groups consisting of heterogeneous agents maintain *stable* cooperation. The bad news is that, unlike in homogeneous groups, the two instruments might not permit groups to maximize group earnings when this implies a high level of inequality. Concerns for this inequality appear to loom large in our experiment, despite the use of a real-effort task designed to make inequality more acceptable (by assigning higher benefits from cooperation

to those who performed better in the task) and the existence of a coercive instrument (punishment) to enforce it. While a few groups do agree to follow contribution rules maximizing group earnings in *ComEarly*, the vast majority agrees on rules which either prioritize equality over efficiency concerns or strike a compromise between the two.

Our findings suggest that in some instances in which cooperation is required, the main problem for mechanism design may not be the tension between private and public interest; after all, individuals stick to agreements with considerable positive contributions. Instead, the main problem may be agents' concerns for equality, which can prevent the emergence of efficient cooperation without the provision of additional incentives. Mechanisms which reduce the tension between equality and efficiency/equity may be the most successful. Of course, like with any empirical finding, generalizations should be made with care as certain forces that may be important in the field have been neutralized in our study. For instance, perfect monitoring is costless in our setting, which is clearly a simplifying assumption. The temptation to violate a covenant will likely be higher in environments with costly monitoring. At the same time, the world outside the laboratory is characterized by diversity in multiple dimensions (e.g., wealth, nationality, religion, educational background, age), increasing the potential for disagreement regarding preferred outcome profiles. Many types of heterogeneity have the potential to create normative conflict, perhaps making it more difficult to reach covenants. Future research could usefully investigate factors that influence whether ex-post normative conflict is resolvable or not.

References

- Andrighetto, G., J. Brandts, R. Conte, J. Sabater-Mir, and H. Solaz (2013). Punish and voice: Punishment enhances cooperation when combined with norm-signalling. *PLoS ONE* 8(6).
- Balafoutas, L., R. Kerschbamer, and M. Sutter (2012). Distributional preferences and competitive behavior. *Journal of Economic Behavior & Organization* 83, 125-135.
- Bland, J., and N. Nikiforakis (2015). Coordination with third-party externalities. *European Economic Review* 80, 1-15.
- Bochet, O., T. Page, and L. Putterman (2006). Communication and punishment in voluntary contribution experiments. *Journal of Economic Behavior and Organization* 60, 11-26.
- Bochet, O., and L. Putterman (2009). Not just babble: A voluntary contribution experiment with iterative numerical messages. *European Economic Review* 3, 309-326.
- Bock, O., I. Baetge, and A. Nicklisch (2014). hroot: Hamburg registration and organization online tool. *European Economic Review* 71, 117-120.

- Brosig, J., J. Weimann, and A. Ockenfels (2003). The effect of communication media on cooperation. *German Economic Review* 4(2), 217–241.
- Buckley, E., and R. Croson (2006). Income and wealth heterogeneity in the voluntary provision of linear public goods. *Journal of Public Economics* 90(4-5), 935-955.
- Cabrales, A., R. Miniaci, M. Piovesan, and G. Ponti (2010). Social preferences and strategic uncertainty: An experiment on markets and contracts. *American Economic Review* 100, 2261-2278.
- Cappelen, A.W., A. Drange Hole, A., E.Ø. Sørensen, and B. Tungodden, B. (2007). The pluralism of fairness ideals: An experimental approach. *American Economic Review* 97(3), 818-827.
- Cason, T., and L. Gangadharan (2015). Promoting cooperation in nonlinear social dilemmas through peer punishment, *Experimental Economics* 18, 66–88.
- Cason, T., and L. Gangadharan (2016). Swords without covenants do not lead to self-governance, *Journal of Theoretical Politics* 28(1), 44-73.
- Charness, G., and M. Rabin (2002). Understanding social preferences with simple tests. *The Quarterly Journal of Economics* 117(3), 817-869.
- Chmura, T., S. Kube, T. Pitz, and C. Puppe (2005). Testing (beliefs about) social preferences: Evidence from an experimental coordination game. *Economics Letters* 88(2), 214-220.
- Dekel, S., S. Fischer, and R. Zultan (2017). Potential Pareto public goods. *Journal of Public Economics*, 146(c), 87-96.
- Denant-Boemont L., D. Masclet, and C. Noussair (2007). Punishment, counterpunishment and sanction enforcement in a social dilemma experiment. *Economic Theory* 33, 145-167.
- Dickinson, D.L., and R.M. Isaac (1998). Absolute and relative rewards for individuals in team production. *Managerial and Decision Economics* 19, 299-310.
- Dickinson, D.L. (2001). The carrot vs. the stick in work team motivation. *Experimental Economics* 4(1), 107-124.
- Egas, M., and A. Riedl. 2008. The economics of altruistic punishment and the maintenance of cooperation. *Proceedings of the Royal Society B* 275, 871–78.
- Engelmann, D. and N. Nikiforakis (2015). In the long run we are all dead: On the benefits of peer punishment in rich environments. *Social Choice and Welfare* 45 (3), 561-577.
- Engelmann, D. and M. Strobel (2004). Inequality aversion, efficiency, and maximin preferences in simple distribution experiments. *American Economic Review* 94(4), 857-869.
- Erkal, N., L. Gangadharan, and N. Nikiforakis (2011). Relative earnings and giving in a real-effort experiment. *American Economic Review* 101(7), 3330-3348.
- Faravelli, M., O. Kirchkamp, and H. Rainer (2013). The effect of power imbalances on incentives to make non-contractible investments. *European Economic Review* 61, 169-185

- Fehr, E., and S. Gächter (2000). Cooperation and punishment in public goods experiments. *American Economic Review* 90 (4), 980-994.
- Fehr, E., and S. Gächter (2002). Altruistic punishment in humans. *Nature* 415, 137-140.
- Fehr, E. and K. Schmidt (1999). A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics* 114(3), 817-868.
- Fischbacher, U. (2007). z-Tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics* 10(2), 171-178.
- Fisher, J., R. Isaac, J. Shatzberg, and J. Walker (1995). Heterogeneous demand for public goods: behavior in the voluntary contributions mechanism. *Public Choice* 85(3), 249-266.
- Fisman, R., S. Kariv, and D. Markovits (2007). Individual preferences for giving. *American Economic Review* 97(5), 1858-1876.
- Gangadharan, L., N. Nikiforakis, and M.C. Villeval (2017) Normative conflict and the limits of self-governance in heterogeneous populations. *European Economic Review* 100, 143-156.
- Gee, L.K., M. Migueis, and S. Parsa (2017). Redistributive choices and increasing income inequality: experimental evidence for income as a signal of deservingness. *Experimental Economics* 20(4), 894-923.
- Gneezy, U., A. Kajackaite, J. Sobel (2018). Lying aversion and the size of the lie. *American Economic Review* 108 (2), 419-453.
- Hobbes, Thomas. 1960. *Leviathan*. Oxford: Basil Blackwell.
- Isaac, R. M., and J.M. Walker (1988). Communication and free-riding behavior: the voluntary contribution mechanism. *Economic Inquiry* 26, 585-608.
- Janssen, M., R. Holahan, A. Lee, and E. Ostrom (2010). Lab experiments for the study of social-ecological systems. *Science* 328 (5978), 613–17.
- Lundquist, T., T. Ellingsen, E. Gribbe, and M. Johannesson (2009). The aversion to lying. *Journal of Economic Behavior & Organization* 70 (1-2), 81-92.
- Masclet, D., C.N. Noussair, S. Tucker, and M.-C. Villeval (2003). Monetary and nonmonetary punishment in the voluntary contributions mechanism. *American Economic Review* 93(1), 366-380.
- 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. (2010). Feedback, punishment and cooperation in public good experiments. *Games and Economic Behavior* 68, 689-702.
- Nikiforakis, N., and H. Mitchell (2014). Mixing the carrots with the sticks: Third party punishment and reward" *Experimental Economics* 17 (1), 1-23.

- Nikiforakis, N., and H. T. Normann (2008). A comparative statics analysis of punishment in public-good experiments. *Experimental Economics* 11, 358–69.
- Nikiforakis, N., C.N. Noussair, and T. Wilkening (2012). Normative conflict and feuds: The limits of self-enforcement. *Journal of Public Economics* 96 (9-10), 797-807.
- Nikiforakis, N., and D. Engelmann (2011). Altruistic punishment and the threat of feuds. *Journal of Economic Behavior and Organization* 78(3), 319-322.
- Noussair, C.N., and F. Tan (2011). Voting on punishment systems within a heterogeneous group. *Journal of Public Economic Theory* 13(5), 661-693.
- Noussair, C.N., and S. Tucker (2005). Combining monetary and social sanctions to promote cooperation. *Economic Inquiry* 43(3), 649-660.
- Ostrom, E., J. Walker, and R. Gardner (1992). Covenants with and without a sword: self-governance is possible. *American Political Science Review* 86(2), 404-417.
- Regan, P.M., and A. C. Stam. (2000). In the nick of time: conflict management, mediation timing, and the duration of interstate disputes. *International Studies Quarterly* 44 (2), 239–260.
- Reuben, E., and A. Riedl (2013). Enforcement of contribution norms in public good games with heterogeneous populations. *Games and Economic Behavior* 77, 122-137.
- Reuben, E., and F. van Winden (2008). Social ties and coordination on negative reciprocity: The role of affect. *Journal of Public Economics* 92(1-2), 34-53.
- Sutter, M., S. Haiger, and M. Kocher (2010). Choosing the stick or the carrot? – Endogenous institutional choice in social dilemma situations. *Review of Economic Studies* 77(4), 1540-1566.
- Tan, F. (2008). Punishment in a linear public good game with productivity heterogeneity, *De Economist* 156(3), 269-293.
- Yamagishi, T. (1986). The provision of a sanctioning system as a public good. *Journal of Personality and Social Psychology* 51 (1), 110-116.
- Weng, Q., and F. Carlsson (2015). Cooperation in teams: The role of identity, punishment, and endowment distribution. *Journal of Public Economics* 126, 25-38.
- Xiao, E., and D. Houser (2005). Emotion expression in human punishment behavior. *PNAS* 102 (20), 7398-7401.

Online Supplementary Material

Covenants before the swords:
The limits to efficient cooperation in heterogeneous groups

Christian Koch, Nikos Nikiforakis and Charles N. Noussair

Contents

<u>APPENDIX A – ADDITIONAL TABLE AND FIGURES</u>	<u>33</u>
<u>APPENDIX B – THEORETICAL RESULTS</u>	<u>36</u>
B.1 BEHAVIORAL MOTIVES AND TYPES OF NORMATIVE CONFLICT	36
B.2 COMMUNICATION WITHOUT PUNISHMENT	41
B.3 COMMUNICATION WITH PUNISHMENT	47
B.4 PUNISHMENT WITHOUT COMMUNICATION	65
B.5 MISCELLANEOUS RESULTS	79
<u>APPENDIX C – COMMUNICATION DATA.....</u>	<u>83</u>
<u>APPENDIX D – ANALYSIS OF COORDINATION PROBLEMS.....</u>	<u>86</u>
<u>APPENDIX E – ANALYSIS OF VARIANCE OF EARNINGS WITH GROUPS.....</u>	<u>88</u>
<u>APPENDIX F – INSTRUCTIONS.....</u>	<u>89</u>
<u>APPENDIX G – INSTRUCTIONS FOR COMMUNICATION CODERS</u>	<u>106</u>
<u>APPENDIX H – REFERENCES</u>	<u>112</u>

Appendix A – Additional Table and Figures

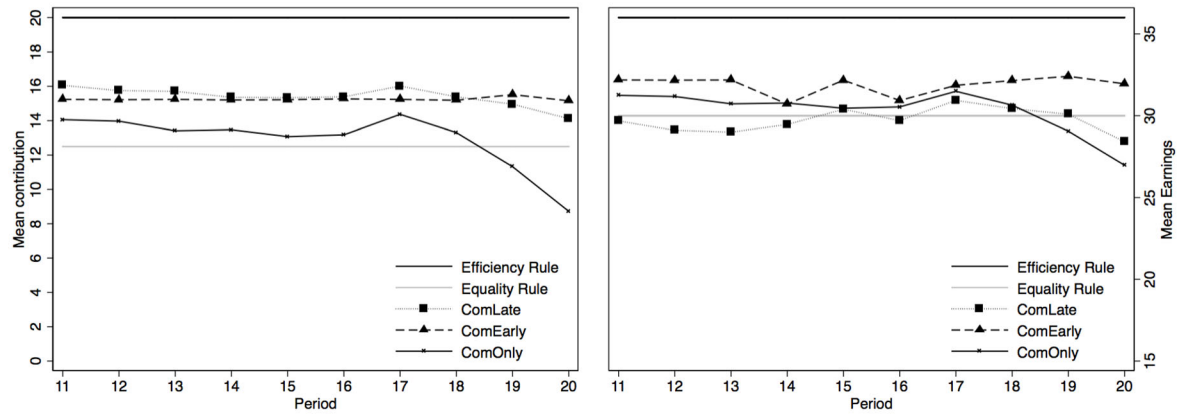


Figure A1 – Average contributions (left) and earnings (right) in Part 2, by Treatment

Table A1– Average Contributions (a), Earnings (b), and Within-group Variance of Earnings (c), by Session Part and Treatment

(a) Contributions			
	All	Part 1	Part 2
<i>Treatments</i>			
<i>ComLate</i>	13.67 (0.92)	11.94 (1.23)	15.40 (0.81)
<i>ComEarly</i>	15.21 (0.66)	15.19 (0.70)	15.24 (0.77)
<i>ComOnly</i>	12.07 (0.95)	11.26 (1.13)	12.88 (0.88)
<i>Statistical Comparison (Mann-Whitney, two tailed): p-values</i>			
<i>ComLate vs. ComEarly</i>	0.207	0.032	0.889
<i>ComLate vs. ComOnly</i>	0.381	0.791	0.041
<i>ComEarly vs. ComOnly</i>	0.014	0.012	0.043
<i>Notes: Standard errors in parentheses.</i>			
(b) Earnings			
	All	Part 1	Part 2
<i>Treatments</i>			
<i>ComLate</i>	27.03 (1.92)	24.34 (2.06)	29.72 (1.95)
<i>ComEarly</i>	31.66 (0.48)	31.45 (0.85)	31.87 (0.63)
<i>ComOnly</i>	29.66 (0.75)	29.01 (0.90)	30.31 (0.71)
<i>Statistical Comparison (Mann-Whitney, two tailed): p-values</i>			
<i>ComLate vs. ComEarly</i>	0.014	0.006	0.414
<i>ComLate vs. ComOnly</i>	0.261	0.098	0.398
<i>ComEarly vs. ComOnly</i>	0.036	0.040	0.093
<i>Notes: Standard errors in parentheses.</i>			
(c) Within-group Variance of Earnings			
	All	Part 1	Part 2
<i>Treatments</i>			
<i>ComLate</i>	268.9 (31.1)	290.9 (32.6)	246.8 (41.6)
<i>ComEarly</i>	181.3 (40.3)	197.9 (41.5)	164.8 (42.3)
<i>ComOnly</i>	124.2 (17.0)	138.2 (16.6)	110.1 (19.3)
<i>Statistical Comparison (Mann-Whitney, two tailed): p-values</i>			
<i>ComLate vs. ComEarly</i>	0.014	0.072	0.138
<i>ComLate vs. ComOnly</i>	0.261	0.001	0.027
<i>ComEarly vs. ComOnly</i>	0.036	0.542	0.842
<i>Notes: Standard errors in parentheses.</i>			

Table A2– Determinants of Sanctioning

	(1)	(2)	(3)	(4)	(5)
	Sanctions	Sanctions	Sanctions	Sanctions	Sanctions
ComEarly	-0.069*** (0.013)	-0.058*** (0.013)	-0.050*** (0.012)	-0.050*** (0.012)	-0.050*** (0.012)
Covenant	-0.048*** (0.005)	-0.035*** (0.005)	-0.032*** (0.004)	-0.032*** (0.004)	-0.032*** (0.004)
Group cont.		-0.002*** (0.000)	-0.001*** (0.000)	-0.001* (0.000)	-0.000 (0.000)
Abu Dhabi		-0.011 (0.014)	-0.002 (0.012)	-0.002 (0.012)	-0.002 (0.012)
Punished prior period			0.008*** (0.001)	0.008*** (0.001)	0.008*** (0.001)
Period			-0.003*** (0.000)	-0.003*** (0.000)	-0.003*** (0.000)
Punisher cont.				0.001 (0.001)	0.001 (0.001)
Target cont.				-0.003*** (0.000)	-0.004*** (0.001)
Punisher- High type					-0.018 (0.014)
Target - High type					0.024*** (0.006)
Constant	0.101*** (0.009)	0.196*** (0.015)	0.172*** (0.014)	0.172*** (0.014)	0.169*** (0.016)
Observations	8160	8160	7752	7752	7752
R-squared	0.051	0.052	0.096	0.099	0.103

Notes: Random-effect panel regression. The dependent variable is the quantity of sanctions that one player assigns to another. In other words, it is the number of punishment points, over all punishment stages, that player i directs to player j in a period. Random effects are at the subject level. Very similar results emerge when random effects are specified at the group level. Standard errors in parenthesis.
 *Significant at 10%; ** significant at 5%; *** significant at 1%.

Appendix B – Theoretical results

This appendix provides formal propositions (and their proofs) for the intuitive insights offered in the main text. The aim of our simple theoretical framework is to help the reader to appreciate the conditions under which heterogeneous benefits from cooperation can lead to normative conflict and, more importantly, when the latter will be an obstacle to efficient cooperation. Our analysis is based on three behavioural types introduced in Section 3.

B.1 Behavioral motives and types of normative conflict

We focus our analysis on the case of *heterogeneous* returns from the public good, as this is the environment subjects face in the experiment. Let us, however, start by discussing the case of *homogenous* returns first to be able to explore the implications of heterogeneous returns as a second step. For most parts of this appendix, we will consider a more general environment than that of the experiment. We assume first that n players that all receive the same benefit m from the public good.

Selfish money maximizers are assumed to have preferences given by $u_i(\pi) = \pi_i$. As discussed in the main text, we follow Fehr and Schmidt (1999) in modelling the preferences of those who dislike inequality. The utility of such an individual i is given by:

$$u_i(\pi) = \pi_i - \alpha_i \frac{1}{n-1} \sum_{j \neq i} \max(\pi_j - \pi_i, 0) - \beta_i \frac{1}{n-1} \sum_{j \neq i} \max(\pi_i - \pi_j, 0), \quad (\text{B1})$$

where α_i measures the disutility i derives from *disadvantageous* inequality and β_i the disutility s/he derives from *advantageous* inequality, with $\alpha_i \geq \beta_i > 0$, and $\beta_i \leq 1$. In addition, the payoff (after the contribution stage) is given by $\pi_i = 20 - c_i + m \sum_{j=1}^n c_j$, with $m < 1$. Naturally, in the homogenous setting, inequality and inequity-averse preferences coincide with each other as all player receive the same return m from the public account.

If we denote player i 's contribution by c_i and the total contribution profile by $\mathbf{c} = c_1 + c_2 + \dots + c_n$, we can rewrite the utility agents derive after the contribution stage as:

$$u_i(\mathbf{c}) = 20 - c_i + m \sum_{j=1}^n c_j - \alpha_i \frac{1}{n-1} \sum_{j: c_i > c_j} \max(c_i - c_j, 0) - \beta_i \frac{1}{n-1} \sum_{j: c_i < c_j} \max(c_j - c_i, 0).$$

We can see that differences in contributions translate easily to differences in payoffs in the homogenous case. The underlying intuition is the following: Let us consider the case that player i deviates from his or her agreed upon contribution of c_i by providing Δ less. In this case, $\pi_i = 20 - (c_i - \Delta) + m(\sum_{j \neq i} c_j + c_i - \Delta)$, implying that i 's earnings increase by $(1 - m)\Delta$. At the same time,

$\pi_j = 20 - c_j + m(\sum_{j \neq i} c_j + c_i - \Delta)$, implying that j 's earnings decrease by $m \Delta$. In other words, if i contributes Δ less, s/he will increase the payoff difference to all other players j by exactly Δ . Assuming that there either was no inequality before the deviation or that player i already had the highest earnings in the group, this increase in payoff differences leads to a total utility cost to other group members of $\frac{n-1}{n-1} \beta_i \Delta$. Intuitively, and as will be shown formally below, these properties imply that homogenous groups will be able to fully cooperate – even without any punishment opportunities – if all group members are sufficiently inequality-averse, i.e. if the cost of deviating due to the implied inequality is higher for them than the monetary benefit: $\beta_i \geq (1 - m)$.

What are the implications of *heterogeneous benefits* in our setting? Let us consider the general case that we have $n_{h(igh)}$ players receiving a high return and $n_{l(ow)}$ players receiving a low return with $m_h > m_l$ and $n_h + n_l = n$. As in our experimental design, we will also assume that $0 < m_l < m_h < 1$. Notably, we will also assume that the joint surplus of high-return players is greater than one, i.e., $n_h \cdot m_h > 1$. We will, however, only need this assumption in some parts of this appendix. Let $\mathbf{c} = \{c_1, c_2, \dots, c_n\}$ again be a group's total contribution vector. Consider the case that a player i who receives a high return from the public good deviates from his or her contribution of c_i by providing Δ less, as in the previous paragraph. In this case, $\pi_i = 20 - (c_i - \Delta) + m_h(\sum_{j \neq i} c_j + c_i - \Delta)$, implying that i 's contributions ceteris paribus increases his or her payoff by $(1 - m_h)\Delta$. At the same time, $\pi_j = 20 - c_j + m_j(\sum_{j \neq i} c_j + c_i - \Delta)$, implying that j 's payoff decreases by $m_h \Delta$ if j is a high type and $m_l \Delta$ if j is a low type. Similarly, a deviation of a low-return player i , ceteris paribus, increases his or her own payoff by $(1 - m_l)\Delta$, while it decreases j 's payoffs – depending on the type – as before. Thus, unlike in the homogenous case, the utility cost of deviating due to increased inequality becomes dependent on which players are matched. For players with the same benefit the payoff difference is simply Δ , as in the homogeneous case.

Let us denote the difference in returns as $\hat{m} = m_h - m_l$. We can see that a Δ deviation of a low-return player inflicts a payoff difference of $(1 + \hat{m})\Delta$ on high-return players. High-return players, on the other hand, only inflict a payoff difference of $(1 - \hat{m})\Delta$ on a low-return player. These changes imply that heterogeneous groups will still be able to cooperate even without any punishment opportunities if all group members are sufficiently inequality-averse, as in a case with homogenous payoffs. Cooperation is possible if the cost of deviating due to the rise in inequality is greater than the monetary cost. Under heterogeneity, the precise conditions are somewhat more complex, reflecting the aforementioned differences. Below, proposition 1 will outline these conditions: $\frac{n_h - 1 + n_l(1 - \hat{m})}{n - 1} \beta_i \geq 1 - m_h$ for high-return players and $\frac{n_h(1 + \hat{m}) + n_l - 1}{n - 1} \beta_i \geq 1 - m_l$ for low-return players. In contrast to the homogenous case, groups that consist of only inequality-averse

players will, however, not be able to cooperate efficiently (i.e. each player contributes fully) because their equality benchmark only allows them to agree on the equality rule, as discussed in more detail below.

When agents derive different benefits from each other, equity might replace equality as a benchmark for some agents, i.e., those agents might believe that someone who has a higher return from the public good should also earn more. We assume that inequity-averse individuals have a similar utility function as inequality-averse individuals but their reference payoff is weighed to reflect equity considerations. In particular, we assume that the utility of such an individual is given by:

$$u_i(\pi) = \pi_i - \gamma_i \frac{1}{n-1} \frac{1}{m_j} \sum_{j \neq i} \max(m_i \pi_j - m_j \pi_i, 0) - \delta_i \frac{1}{n-1} \frac{1}{m_i} \sum_{j \neq i} \max(m_j \pi_i - m_i \pi_j, 0), \quad (\text{B2})$$

where $\gamma_i \geq \delta_i > 0$, and that $\delta_i \leq 1$. This modelling implies that someone who has, e.g., twice as high a return from the public good should earn twice as much. If $m_i > m_j$ then i will suffer a loss in utility if her monetary payoff is not sufficiently larger than that of j . The underlying idea of the utility specification is that only when $m_i \pi_j - m_j \pi_i = 0$ is there no utility loss. In other words, inequity is zero if $m_i/m_j = \pi_i/\pi_j$.

Notably, after weighting payoffs according to the returns from the public account, a normalization is applied by dividing any difference in weighted payoffs by m_i or m_j , respectively. The normalization allows γ and δ to have a similar magnitude to the α, β 's of the traditional model of FS. The normalization is based on the following idea: First, if a high benefit player has earnings greater than the equity benchmark, then the payoff advantage in excess of equity towards a low-return player has to be m_h/m_l times the payoff advantage towards a fellow high-return player to imply the same level of advantageous inequity. This is true because in this case our normalization implies that π_h is effectively weighted by m_l/m_h and π_l by 1 (i.e., $\delta_i \frac{1}{m_h} (m_l \pi_h - m_h \pi_l) = \delta_i (\frac{m_l}{m_h} \pi_h - \pi_l)$). Second, if a high-return player has lower earnings than the equity benchmark, then a low-return player's payoff advantage in excess of equity must be m_l/m_h times the one of a fellow high-return player to imply the same level of disadvantageous inequity. Similarly as before, π_l is effectively weighted by m_h/m_l and π_h by 1 in this case (i.e., $\gamma_i \frac{1}{m_l} (m_h \pi_l - m_l \pi_h) = \gamma_i (\frac{m_h}{m_l} \pi_l - \pi_h)$). Similar arguments apply for a low-return player being ahead or behind the equity benchmark. The chosen normalization seems to reflect most closely the idea that inequity-averse players treat high-return and low-return players differently according to their returns. More generally, let us point out that other conceptualizations of aversion against inequity are possible, but we think that

our particular functional form captures that idea that high types might be guided by the payoff advantage they earned in the real-effort task. Of course, if everyone contributes fully, no inequity arises and efficiency is reached.

In our simple environment, equity-minded individuals pursue efficiency (i.e. efficient outcomes) for the following reason. If one accepts that participants may have unequal earnings, i.e. their earnings follow the returns of the public good, no “redistribution” in terms of lower contributions for low-return players is necessary. When all subjects contribute fully, all players (including low-return players) get what they “deserve”. In this sense the efficient outcome is an equitable outcome which equity-minded players favor. Of course, another way of modelling a preference for the efficient outcome would be to simply assume that agents, and in particular high-return players, directly care about efficiency. Notably, such a simple approach would not be able to explain why players punish each other and engage in feuds. The latter types of behavior reduce (and do not directly increase) efficiency. Of course, a more complex utility function may not only feature a direct efficiency concern, but also a reciprocity component, in which the benchmark behavior are efficient contributions. Such a hybrid utility function would give rise both to a desire for efficiency and a willingness to punish deviations from efficiency. Put differently, high-return players could – in a complex way – either care for efficiency per se or they could feel entitled – potentially due to self-serving biases – to their advantage in returns that they earned in a real-effort task. We deliberately designed our experiment in a way so that these genuine “efficiency” concerns cannot be distinguished from those that arise from a self-serving bias, as both factors give rise to an inclination towards efficiency (i.e. efficient outcomes).²⁸ We choose to model high-return players as inequity averse as it is a simple and an elegant analog to inequality aversion. We would expect, however, that a more complicated modelling of efficiency concerns (as sketched above) would lead to broadly similar results.²⁹

What are the implications of inequity aversion in our environment? Let $\mathbf{c} = \{c_1, c_2, \dots, c_n\}$ be a contribution profile. If either a high or a low-return player i deviates from this profile, the same implications for her own and other players’ *payoffs* emerge as discussed in the context of inequality

²⁸ In a more complex environment, in which some players’ earnings might be subject to a random negative shock, equity and efficiency concerns may actually lead to different outcomes. In such a case, equity concerns may imply that some form of “redistribution”, in terms of lower contributions, for the player hit by the negative shock is desirable (as final earnings should still follow the return structure of the public good). With efficiency concerns, such redistributions can never be desirable as they would reduce efficiency.

²⁹ Another advantage of our inequity approach is that it seems somewhat more compatible with respect to contributions below the Pareto frontier. Although these contributions are not the focus of our paper, equal and equitable outcomes can arise below the frontier. The efficient outcome is, of course, on the frontier. This implies that *equality* and *equity* concerns (in contrast to *efficiency* concerns) are somewhat more compatible from a modelling point of view.

aversion. Note, however, that the new utility function potentially amplifies the externalities from changes in contributions in utility terms. A player i 's deviation of Δ implies a payoff difference of Δ for player j with the same benefit from the public account. Assuming that there either was no inequality before the deviation or that player i already had higher earnings than player j , player i 's disutility due to advantageous inequity changes by $\frac{\delta_i}{n-1} \frac{1}{m_l} (m_l(1 - m_h) + m_l m_h) \Delta = \frac{\delta_i}{n-1} \frac{m_l}{m_l} (1 - m_h + m_h) \Delta = \frac{\delta_i}{n-1} \Delta$, as expected. However, a deviation of Δ by high-return player i implies a payoff difference of only $(1 - \hat{m})\Delta$ towards low-return player k , as outlined earlier. In utility terms, the change of advantageous inequality is: $\frac{\delta_i}{n-1} \frac{1}{m_h} (m_l(1 - m_h) + m_h m_l) \Delta = \frac{\delta_i}{n-1} \frac{m_l}{m_h} \Delta$. Thus, applying the weights decreases the difference (i.e. $(1 - \hat{m}) < \frac{m_l}{m_h}$).³⁰ The intuition is that in our formalization of inequity, being ahead of a low-return player by one unit is less hurtful than being ahead of a high-return player by one unit, a reflection of the fact that different players have different claims with respect to public account. Similarly, a deviation by a low-return player will lead to a change in inequity of $\frac{\delta_i}{n-1} \Delta$ with respect to another low and $\frac{\delta_i}{n-1} \frac{m_h}{m_l} \Delta$ with respect to a high-return player.

Notably, groups that only consist of inequity-averse players will still be able to cooperate even without punishment if all members are sufficiently inequity-averse. Proposition 1 below will provide the precise conditions that enable cooperation: $\frac{(n_h-1)+n_l \frac{m_l}{m_h}}{n-1} \delta_i \geq 1 - m_h$ for high-return players and if $\frac{n_h \frac{m_h}{m_l} + (n_l-1)}{n-1} \delta_i \geq 1 - m_l$ for low-return players. This time, of course, players can agree on the efficiency rule, as it is compatible with their payoff benchmark of equity. The equality rule, however, is not reachable any longer.

As already assumed implicitly so far, we are only considering a one-shot public good environment in our formal analysis, i.e., we are focusing on stage-game equilibria and do not model the dynamic aspect of our game. The underlying idea is the following: Our core interest is to get an idea what the impact of our instruments (communication alone, punishment alone and both together) is on cooperation in an environment in which subjects likely have conflicting normative views. Should we expect communication to have a positive effect in such a setting? Would punishment help? Do we need both? In other words, to get a general understanding what might happen when participants disagree about what the right behavioral rule might be in our different treatments, a complete analysis of our rather complex dynamic game structure does not seem necessary (in particular as such an analysis might also be infeasible). For this reason, we focus on

³⁰ Note that $m_h = m_l + \hat{m}$. Thus, we want to show that $1 - \hat{m} > m_l / (m_l + \hat{m})$. We can rearrange: $m_l + \hat{m} - \hat{m}(m_l + \hat{m}) > m_l$. As $(m_l + \hat{m})m_h < 1$, the condition is true.

how the conflicting behavioral forces (behind normative conflict) play out in the stage game. Notably, we find empirical support for the fact that stage-game equilibria seem to be relevant insofar as those groups that seem to be in equilibrium largely do not change their play across communication rounds.³¹

Overall, these arguments suggest that we should expect at least somewhat similar outcomes in a public good setting with heterogeneous benefits as in a setting with homogenous returns (which is discussed in section B.5 of this appendix), as long as we assume that *all* group members are either inequality-averse or inequity-averse. Of course, there are reasons to believe that all group members may not be identical and that we may observe what has been called *normative conflict*. As outlined in the main text, we define *ex-post normative conflict* (EPNC) as a situation in which both inequity- and inequality-averse types coexist in a group. We distinguish between situations of *resolvable ex-post normative conflict* (R-EPNC) which are instances when EPNC does not lead to the unravelling of cooperation and *unresolvable ex-post normative conflict* (U-EPNC) in which EPNC leads to minimal cooperation, a feud or a total destruction of payoffs. In addition, we define *ex-ante normative conflict* (EANC) as a situation in which there is uncertainty about the distributional preferences of other group members. Finally, we will say that there is *no normative conflict* when it is common knowledge that inequity and inequality-averse types do not coexist in a group.

B.2 Communication without punishment

The potential coexistence of behavioral types in any given group implies that individuals are likely to be uncertain about the preferences of others in their group. Communication can serve to reduce this uncertainty and, for simplicity, we will assume that communication fully removes the

³¹ First of all, considering the case in which both communication and punishment is available to groups, most groups form a covenant in *ComEarly* (Part 1)/ *ComLate* (Part 2) during the communication stage (>80%). Notably, most of these groups follow through on this covenant during the subsequent three rounds after the communication stage (>80%), suggesting that they seem to be in equilibrium. Looking at these groups, most of them do not change the rule they implement between communication stages 1 and 2 (*ComEarly*: >80%, *ComLate*: >90%) or between stages 2 and 3 (*ComEarly*: 78%, *ComLate*: 90%). This seems to support the relevance of stage game equilibria (insofar as e.g. there is no renegotiation). When only punishment is available, most groups (87%) do not follow a consistent pattern over three rounds and there is lots of punishment, suggesting that most groups are actually not in equilibrium. When only communication is available, many groups form covenants (>80%) and those who follow through on them usually also do not change the rule they follow during the next communication period (>80%). Of course, more groups do not adhere to their covenant as no disciplining device is available.

uncertainty characterizing ex-ante normative conflict.³² Thus, our analysis for the environment in which our subjects can only communicate but not punish assumes complete information about preferences. While this assumption is, of course, often made even in the absence of communication in the analysis of a homogenous public-good setting, it seems much less straightforward in a setting in which heterogeneity may lead to normative conflict, at least without any form of communication.

Our first proposition distinguishes between two cases: no normative conflict vs. normative conflict. It shows that in former case, cooperation can be obtained in equilibrium if *all* players are either (i) *sufficiently inequality-averse* or (ii) *sufficiently inequity-averse*. This condition implies that in the presence of *one* selfish player ($\alpha_i = \beta_i = \gamma_i = \delta_i = 0$), cooperation cannot be maintained. In the latter case, i.e. when inequity and inequality averse agents coexist, we assume that high-return players are inequity-averse and low-return players are inequality-averse (and discuss the reverse case below in B.5). Notably, with normative conflict, we show that cooperation unravels. Notably, the proposition refers to *symmetric* and *equity-symmetric* payoff vectors. A symmetric payoff vector is characterized by the fact that all players receive identical payoffs. An equity-symmetric payoff vector is characterized by the fact that all high-return players receive the same payoff, π_h and all low-return players receive π_l respectively, and it holds that $\pi_h = \frac{m_h}{m_l} \pi_l$.

Proposition 1: (a) [No Normative Conflict]

(i) Suppose that all members are either inequality-averse or selfish (a selfish player has $\beta_i = \alpha_i = 0$). If $\beta_i \leq \frac{n-1}{(n_h)(1+\hat{m})+n_l-3} (1 - m_l)$ for all low-return and $\beta_i \leq \frac{n-1}{(n_l-2)(1-\hat{m})+n_h-1} (1 - m_h)$ for all high-return players, then

1. there are no equilibria with asymmetric payoffs,
2. the vector of zero contributions is a Nash equilibrium, and
 - a. if $\frac{n_h-1+n_l(1-\hat{m})}{n-1} \beta_i < 1 - m_h$ for some high-return or if $\frac{n_h(1+\hat{m})+n_l-1}{n-1} \beta_i < 1 - m_l$ for some low-return players, then there are no other equilibria,
 - b. if $\frac{n_h-1+n_l(1-\hat{m})}{n-1} \beta_i \geq 1 - m_h$ for all high-return and if $\frac{n_h(1+\hat{m})+n_l-1}{n-1} \beta_i \geq 1 - m_l$ for all low-return players, then any vector of contributions that leads to equal payoffs for all players, including the **equality rule**, is a Nash equilibrium.

³² This is clearly a strong assumption. There is, of course, ample evidence showing that many individuals incur costs when lying (Lundquist et al. 2009, Gneezy et al. 2018). Furthermore, in our setting with fixed IDs, individuals have an incentive to maintain a reputation of being perceived as honest. We recognize, however, that some players might still be willing to lie, especially in the absence of punishment opportunities. Nevertheless, we abstract from this scenario in our analysis for tractability, since the possibility of some players lying only reduces the predicted efficacy of communication, but does not affect the comparative statics in our analysis, as will become apparent below.

(ii) Suppose that all group members are either inequity-averse or selfish. If $\delta_i \leq \frac{n-1}{(n_l-3)+n_h \frac{m_h}{m_l}} (1 - m_l)$ for all low-return players and if $\delta_i \leq \frac{n-1}{(n_h-2)+(n_l-1) \frac{m_l}{m_h}} (1 - m_h)$ for all high-return players,

then

1. there are no equilibria with positive contributions that do not have equity-symmetric payoffs

2. the vector of zero contributions is a Nash equilibrium, and

- a. if $\frac{(n_h-1)+n_l \frac{m_l}{m_h}}{n-1} \delta_i < 1 - m_h$ for some high-return players or if $\frac{n_h \frac{m_h}{m_l} + (n_l-1)}{n-1} \delta_i < 1 - m_l$ for some low-return players, there are no other equilibria,
- b. if $\frac{(n_h-1)+n_l \frac{m_l}{m_h}}{n-1} \delta_i \geq 1 - m_h$ for all high-return players and if $\frac{n_h \frac{m_h}{m_l} + (n_l-1)}{n-1} \delta_i \geq 1 - m_l$ for all low-return players, then any vector of contributions that leads to payoffs in which high types earn $\frac{m_h}{m_l}$ times the amount of low types, including the **efficiency rule**, is a Nash equilibrium.

(b) [Ex-Post Normative Conflict]

Suppose that high-return players are either inequity-averse or selfish and low-return players are either inequality-averse or selfish. If $\delta_i \leq \frac{n-1}{(n_h-2)+(n_l-1) \frac{m_l}{m_h}} (1 - m_h)$ for all high-return individuals

and if $\beta_i \leq \frac{n-1}{(n_h)(1+\hat{m})+n_l-3} (1 - m_l)$ for all low-return individuals, the vector of zero contributions is the only Nash equilibrium.

Proof.

Part (a).

(i) Consider any vector $\mathbf{c}$ that leads to a vector of non-identical payoffs $\boldsymbol{\pi}$. Let i be a high-return player with the lowest payoff resulting from $\mathbf{c}$. Suppose that agent i deviates and contributes slightly less than c_i , say $c_i - \Delta$, where Δ is such that $c_i - \Delta$ is still strictly lower than the second lowest payoff resulting from the contribution vector $\mathbf{c}$. By doing so, i increases his or her own private payoff by $(1 - m_h)\Delta$. In addition, i increases his or her utility by reducing the payoff difference with players with higher earnings. Since $\mathbf{c}$ leads to non-identical payoffs, there is at least one player with higher earnings than i . If this is a low-return player, the minimal gain to i is $\frac{\alpha_i(1-\hat{m})}{n-1} \Delta$ (which is larger or equal than $\frac{\beta_i(1-\hat{m})}{n-1} \Delta$). The minimal gains to a high-return player would exceed $\frac{\alpha_i(1-\hat{m})}{n-1} \Delta$. In addition, however, the deviation is associated with a utility loss as it increases

inequality with people who previously had the same payoff. There are at most $n - 2 [= (n_l - 1) + (n_h - 1)]$ such players, resulting in a maximum loss of $\frac{\beta_i}{n-1} [(n_l - 1)(1 - \hat{m}) + (n_h - 1)]\Delta$. Thus, it will always be beneficial for a (high-return) player with the lowest payoff to deviate as long as

$$\left((1 - m_h) + \frac{\alpha_i(1 - \hat{m})}{n - 1} \right) \Delta > \frac{\beta}{n - 1} [(n_l - 1)(1 - \hat{m}) + (n_h - 1)] \Delta.$$

This condition holds if $\beta_i \leq \frac{n-1}{(n_l-2)(1-\hat{m})+n_h-1} (1 - m_h)$.

Now, suppose the deviating player i is a low-return player: i gains $(1 - m_l)\Delta$ as well as at least $\frac{\alpha_i}{n-1}\Delta$ in inequality reduction (assuming again that one low-return player³³ earns more than i). In addition, there are at most $n - 2 [= (n_l - 2) + (n_h - 1)]$ players with the same initial utility. Thus, deviating is profitable for a high-return player as long as

$$\left((1 - m_l) + \frac{\alpha_i}{n - 1} \right) \Delta > \frac{\beta_i}{n - 1} [(n_l - 2) + n_h(1 + \hat{m})] \Delta.$$

This condition holds if $\beta_i \leq \frac{n-1}{(n_l-3)+n_h(1+\hat{m})} (1 - m_l)$. In case conditions are fulfilled both for low-return and high-return players, no player – under any circumstances – wants to be among the players with lowest payoffs. Therefore, there cannot be an equilibrium in which players have different earnings.

We will now consider symmetric equilibria in which all players earn an equal payoff. Note that the vector of zero contributions, $(0, \dots, 0)$, is a Nash equilibrium. Indeed, in this case, the utility of all agents is simply 20 and any unilateral deviation will lead to lower utility as $m_i < 1$ and, in addition, disadvantageous disutility will arise. Let $\bar{\mathbf{c}}$ be a vector of contributions that leads to a vector of identical payoffs $\boldsymbol{\pi} (= \pi, \dots, \pi)$. A high-return player gains $(1 - m_h)\Delta$ by deviating down by Δ . At the same time, the disutility due to advantageous inequality to the $n_h - 1$ other high-return players and the n_l low-return players increases by $\frac{\beta_i}{n-1} (n_h - 1 + n_l(1 - \hat{m}))\Delta$. Thus, only if the benefit of the deviation is smaller than its cost, $\frac{n_h-1+n_l(1-\hat{m})}{n-1} \beta_i > 1 - m_h$, for all high-return player equilibria with positive contributions can exist. The condition for low-return players is derived similarly.

(ii) As outlined above, we denote a vector of payoffs as *equity-symmetric* high-return players receive the same payoff, π_h and all low-return players receive π_l respectively, and it holds that $\pi_h = \frac{m_h}{m_l} \pi_l$. In addition, we define *equity-adjusted* payoffs $\tilde{\boldsymbol{\pi}}$ as a payoff vector in which low-return

³³ As in the previous case, the gain is minimal with respect to low-return player compared to a high-return one.

players' payoffs are multiplied by $\frac{m_h}{m_l}$ with $\tilde{\pi}_l = \frac{m_h}{m_l} \pi_l$. In other words, a payoff vector that is equity-symmetric implies that the associated vector of equity-adjusted payoffs has identical entries.

Consider any vector $\mathbf{c}$ that leads to a vector of payoffs $\boldsymbol{\pi}$ that are not equity-symmetric. Let i be a high-return player with the lowest of the equity-adjusted payoffs resulting from $\mathbf{c}$. Suppose that agent i deviates and offers slightly less than c_i , say $c_i - \Delta$, where Δ is such that $c_i - \Delta$ results in payoffs that are still strictly lower than the second lowest of the equity-adjusted payoffs resulting from the contribution vector $\mathbf{c}$. By doing so, i increases his or her own private payoff by $(1 - m_i)\Delta$. In addition, i increases his or her utility by reducing the difference to players with (equity-adjusted) higher earnings. Since $\mathbf{c}$ leads to payoffs that are not equity-symmetric, there is at least one such player – either a low or a high-return player. We assume the deviator is a high-return player. By deviating, i gains $(1 - m_h)\Delta$ in payoffs and at the same time the other player loses $m_j\Delta$, implying that the minimal gain in utility is $\frac{\gamma_i}{n-1} \Delta$ (which is larger or equal than $\frac{\delta_i}{n-1} \Delta$). Here, it does not matter whether the gain is with respect to a high or a low-return player. While the gain in payoff is lower with respect to a low-return player, it is more beneficial to gain against low-return players in terms of utility.³⁴ In addition, however, the deviation is associated with a utility loss as it increases inequity with people of the same equity-adjusted payoff. There are at most $n - 2$ such players, resulting in a maximum loss of $\frac{\delta_i}{n-1} \left[(n_h - 1) + (n_l - 1) \frac{m_l}{m_h} \right] \Delta$. Thus, deviating is profitable for high type as long as

$$\left((1 - m_h) + \frac{\gamma_i}{n-1} \right) \Delta \geq \frac{\delta_i}{n-1} \left[(n_h - 1) + (n_l - 1) \frac{m_l}{m_h} \right] \Delta.$$

This condition holds if $\delta_i \leq \frac{n-1}{(n_h-2)+(n_l-1)\frac{m_l}{m_h}} (1 - m_h)$. A similar argument– leads to the condition for the lower types: $\delta_i \leq \frac{n-1}{(n_l-3)+n_h\frac{m_h}{m_l}} (1 - m_l)$. If the two conditions for low-return and high-return players are fulfilled, no player wants to be among the players with lowest payoffs and potentially highest contributions. Therefore, there cannot be an equilibrium with both positive contributions and equity-asymmetric payoffs.

Note that the vector of zero contributions, $(0, \dots, 0)$, is a Nash equilibrium. Indeed, in this case, the utility of all agents is simply 20 and any unilateral deviation will lead to lower utility as $m_i < 1$ and, in addition, disadvantageous inequity will arise. Of course, high types will suffer

³⁴ With respect to a high-return player the payoff difference is reduced by $(1 - m_h + m_h)\Delta = \Delta$. In case of a low-return player, the payoff difference is reduced by $(1 - m_h + m_l)\Delta = (1 - \hat{m})\Delta$. Curically in utility terms the reduction is the same: $\frac{\gamma_i}{n-1} \frac{1}{m_l} [m_h(m_l) + m_l(1 - m_h)]\Delta = \frac{\gamma_i}{n-1} [m_h + (1 - m_h)]\Delta$.

disadvantageous inequity in any case, but they cannot decrease this inequity by means of a unilateral deviation.

Now let $\bar{c}$ be a vector of equity-symmetric contributions, i.e. a vector of payoffs in which high-return players earn $\frac{m_h}{m_l}$ times the amount that low-return players do. A high type gains $(1 - m_h)\Delta$ by deviating and reducing his or her contributions by Δ . At the same time the disutility due to advantageous inequity to the $n_h - 1$ other high types and the n_l low types increases by $\frac{\delta_i}{n-1} \left[(n_h - 1) + n_l \frac{m_l}{m_h} \right] \Delta$. The only difference to the proof of part (i) is that the increase in disutility is weighted by the return parameters. Thus, only if the cost of deviation is larger than its benefit, $\frac{(n_h-1)+n_l \frac{m_l}{m_h}}{n-1} \delta_i > 1 - m_h$, for all high-return players, do equilibria with positive contributions exist. Similarly, a low-return player could gain $(1 - m_l)\Delta$ but faces disutility from n_h high-return players and $n_l - 1$ other low-return players, resulting in $\frac{\delta_i}{n-1} \left[n_h \frac{m_h}{m_l} + (n_l - 1) \right] \Delta$. Thus only if the cost of deviation is larger than its benefit, $\frac{n_h \frac{m_h}{m_l} + (n_l-1)}{n-1} \delta_i > 1 - m_l$, for all low-return players, do equilibria with positive contributions exist.

Part (b).

When everyone contributes zero, the payoff of all agents is simply 20, the endowment. Any unilateral deviation will not be profitable for either low-return or high-return individuals. Since zero contributions result in equal earnings, low-return players will not deviate as it decreases their payoff by $1 - m_l$ and leads to additional inequality. While high-return players experience disutility due to disadvantageous inequity in this equilibrium, unilateral deviations are still not profitable for them, as they decrease their payoff by $1 - m_h$ and even increase the disutility due to inequity.

The conditions are taken directly from part (a) of the proposition. If those conditions are fulfilled (which they are under the parameters of the experiment), high types will deviate from any (strictly positive) contribution schedule that involves payoffs that are not equity-symmetric and low types will deviate from any contribution schedule that involves asymmetric payoffs. Of course, there is no contribution schedule that can lead both to equity-symmetric and symmetric payoffs at the same time. ■

The initial conditions of part (a) of the proposition basically rule out equilibria in which some group members contribute while others free-ride. For the set of parameters used in our experiment,

they are – with one exception³⁵ – always fulfilled (as $\beta, \gamma \leq 1$). Intuitively, with only four players, one selfish player already makes it sufficiently unattractive for the other three players to cooperate. As in the homogenous case (see B.5), groups can cooperate if all group members are either sufficiently inequality-averse or inequity-averse. In the former case, only outcomes that lead to equal earnings, including the *equality rule*, can be rationalized. While in the latter case only outcomes that lead to equitable earnings, including the *efficiency rule*, can be rationalized. In the main text, we focus on the equality and efficiency rules as they generate payoffs along the Pareto-frontier. They are, of course, more attractive than those equilibria with contributions that are not on the Pareto-frontier, i.e. in which high-return players do not contribute fully. Appendix D provides some empirical evidence that players actually do not have problems to coordinate on Pareto-superior outcomes in case they agree on a contribution schedule, at least in case communication is possible. If equity- and equality-minded players coexist within a group, so that there is ex-post normative conflict, cooperation is not possible if players can communicate but not punish. The underlying reason is simple: for cooperation to succeed all group members have to be either sufficiently inequality or inequity-averse, but there is no outcome (with positive contributions) that is both equal and equitable in our setting. In summary, with communication, we may expect some degree of cooperation but only when normative views are aligned. That is why communication alone is unlikely to lead to efficient cooperation in many groups.

B.3 Communication with punishment

While heterogeneity can lead to normative conflict that provides an obstacle for cooperation when players can only communicate, punishment provides a disciplinary instrument that can not only motivate selfish players to contribute but also force players of conflicting normative views to compromise. Of course, on the negative side, if players conflicting views are unresolvable, sanctions will be meted out and can lead to large efficiency losses.

Let us first consider the case of no normative conflict. If all group members are either sufficiently inequality or sufficiently inequity-averse – as in part (a) of the first proposition – the equality and the efficiency rules can also be equilibrium outcomes of the game with punishment. Players simply follow the rule and have neither an incentive to deviate nor to sanction others. As

³⁵ The exception is low-return inequity-averse players. For players with $\delta_i > 0.7$, the relevant condition is not met when using our experimental parameters. For this reason, in principle, equilibria could exist in which one low-return player as well as both high-return players contribute but one low-return players free-rides. Since it would be very costly for the contributing low-return player to be ahead of the high benefit players, s/he still has no incentive to deviate. One can show, however, that even in this situation high-return players still have an incentive to deviate (as $(1 - m_h) - \frac{\delta_i}{3}(1 - \frac{m_L}{m_h}) + \frac{\gamma_i}{3}(1) < 1 - m_h - \frac{1}{3}\frac{m_L}{m_h} > 0$ under the parameter of our experiment). Taking this into account, no other equilibria than the one mentioned in proposition 1 exist in our experimental setting (even if $\delta_i > 0.7$ for low-return players).

the second proposition outlines, the same outcomes can even be obtained in the presence of selfish players. Individuals who care strongly about inequality or inequity will be inclined to sanction others who deviate from the action profile yielding their reference payoff.

Punishment opportunities also fundamentally change the situation when conflicting normative views coexist. While inequity-averse high-return players favor full contributions from everyone, they may be willing to accept slightly lower contributions from low-return players if their aversion is not too strong. Similarly, inequality-averse individuals favor contributions that lead to equal outcomes but they may be willing to contribute slightly more, at least if their aversion to inequality is not strong and they are threatened by sanctions. The following proposition characterizes the conditions under which the efficiency and equality rules, as well as other contribution profiles along the Pareto frontier, can arise in equilibrium.

Proposition 2: (a) [No Normative Conflict]

(i) Suppose there is one high-return player that is sufficiently inequality-averse, i.e., his or her preferences obey $\frac{n_h - 1 + n_l(1 - \hat{m})}{n - 1} \beta_i \geq 1 - m_h$ and $\alpha_i \geq \frac{n - 1}{1 + n_l \hat{m}}$. In addition, suppose that all other players are selfish. Then, the following strategies form a subgame perfect equilibrium:

- In the contribution stage, each player contributes $\bar{c}_i$ such that a vector of equal payoffs $(\pi, \dots, \pi)$ arises. This includes the **equality rule**.
- If each player contributes $\bar{c}_i$, there are no sanctions in the second stage. If a player j deviates by $\bar{c}_j - \Delta, \Delta > 0$, then the sufficiently inequity-averse player i – the enforcer – chooses to assign the following punishment points: $p_{ij} = \Delta(1 + \hat{m}) + 1$ if the deviator is a low-return player, $p_{ij} = \Delta + 1$ if the deviator is a high-return player, $p_{ik} = 1$ for $k \neq j$ if k is high-return and $p_{ik} = \Delta \hat{m} + 1$ if k is low-return. Other players do not sanction.

(ii) Suppose there is one high-return player that is sufficiently inequity-averse, i.e., his or her preferences obey $\frac{(n_h - 1) + n_l \frac{m_l}{m_h}}{n - 1} \delta_i \geq 1 - m_h$ and $\gamma_i \geq n - 1$. In addition, suppose that all other players are selfish, then the following strategies form a subgame perfect equilibrium:

- In the contribution stage, each player contributes $\bar{c}_i$ such that a vector of payoffs emerges in which high-return players earn $\frac{m_h}{m_l}$ times the amount of low-return players. The **efficiency rule** satisfies this condition.
- If each player contributes $\bar{c}_i$, there are no sanctions in the second stage. If a player j deviates by $\bar{c}_j - \Delta$, then the sufficiently inequality-averse player i – the enforcer – chooses to assign the following punishment points: $p_{ij} = \Delta + \frac{m_l}{m_h}$ if the deviator is a low-return

player, $p_{ij} = \Delta + 1$ if the deviator is a high-return player, $p_{ik} = 1$ for $k \neq j$ if k is high-return and $p_{ik} = \frac{m_l}{m_h}$ if k a low-return player. Other players do not punish.

(b) [Ex-Post Normative Conflict]

Let there be at least one (or two) inequity-averse high-return individual(s), and one inequality-averse low-return individual. All other players are selfish. Then, the high-return player's aversion to (disadvantageous) inequity γ_i determines $c^{\min}$, i.e., the minimum level of contributions by low-return players that inequity-averse high-return players are willing to accept (when $c_{\text{high}} = 20$). The low-return player's aversion against (disadvantageous) inequality α_i determines $c^{\max}$, i.e., the maximum level inequality-averse low-return players are willing to contribute under the threat of punishment (when $c_{\text{high}} = 20$). We focus on potential symmetric³⁶ equilibria on the Pareto-Frontier, in which all high types contribute fully while all low types contribute the same amount $\tilde{c}^{\text{eq}} \leq \tilde{c} \leq 20$, where $\tilde{c}^{\text{eq}}$ is the contribution that will lead to equal earnings.

- i. If $c^{\min} \leq c^{\max}$, the contribution schedule in which high-return players contribute fully and low ones contribute $c^{\min}$, while no player punishes in the second stage, is a subgame perfect equilibrium. If $(1 - m_l) + \alpha_i \frac{n_h}{n-1} (1 + \hat{m}) - \beta_i \frac{n_l-1}{n-1} > 0$, it is the only symmetric subgame perfect equilibrium that does not involve sanctions.
- ii. If $c^{\min} > c^{\max}$, there is no equilibrium without sanctioning with an outcome on the Pareto-frontier. If aversions to inequity and inequality are not too high – or more precisely $c^{\min}$ and $c^{\max}$ are not too far apart from each other – either zero contributions or equilibria below the Pareto-frontier (both without sanctioning) can be outcomes of equilibrium play. If $c^{\min}$ and $c^{\max}$ are sufficiently apart from each other so that, $c^{\min} > \frac{\hat{m}(20n_h + n_l c^{\max}) - (20 - c^{\max})}{1 + \hat{m}n_l} + \frac{m_l}{m_h} 20 > c^{\max}$, sanctioning will occur. In case that in addition, aversions – in absolute terms – are sufficiently strong (i.e.: $c^{\min} > 20 - \frac{m_l}{m_h} 20(m_h - m_l) - \frac{m_l}{m_h} + 1$ and $c^{\max} < \frac{20(m_l + 1 - n_h \hat{m}) + 1 - m_l/m_h}{1 + n_l \hat{m}} - a$ total destruction of payoffs is inevitable.³⁷

³⁶ For simplicity, we will call *semi-symmetric* equilibria, in which all high-return players contribute the same amount (e.g.: $c_h = 20$) and all low-return players do the same ($c_l = \tilde{c}$). Note that we also consider symmetric equilibria only, since there is no reason to suppose that individuals of the same type would adopt different strategies from each other.

³⁷ The precise conditions provided here also rest on the regularity condition $\frac{m_l}{m_h} < \frac{1 - \hat{m}n_h}{1 + \hat{m}n_l}$, which is fulfilled under the parameters of our experiment.

Proof.

Before providing the proof of the proposition, we will make some remarks about sanctions, feuds and a potential destruction of payoffs. Note that we only assume *one* punishment stage in our setting, although we obviously have implemented multiple stages in our experiment. In practice, it may make a lot of difference for our participants whether one or more stages of punishment exist. In particular, subjects will only learn after the first punishment stage whether they have been sanctioned or not and may react accordingly, potentially investing in counter-sanctions and making initial sanctions costlier. We first note that such a lack of information about others' behavior does not play a role in an equilibrium analysis, as players know how others plan to behave and to what extent they sanction. In addition, note that we apply a flat fee for punishment so that paying the fee in one-stage allows for costless punishment in all subsequent stages.³⁸ Furthermore, any player is free to punish any other player at any punishment stage.³⁹ We argue next that these features imply that one-stage or multiple-stage punishment will lead to the same outcome.

Consider a vector of sanctions over multiple stages of all group members. Agents could, at least in theory, always deliver the sum of this vector in one period. The difference between having two or more rounds of sanctions is that retaliatory sanctions can be directly applied in one stage or spread over several stages. Does introducing more rounds change the incentives to sanction and counter-sanction? Whether sanctions of player i create an incentive for another group member, player j , to engage in punishment as well depends solely on the overall payoff implications of the sum of i 's sanctions and not on the punishment stage in which those sanctions have been applied, at least in our framework based on distributional preferences. A vector of sanctions over multiple rounds leads to the same incentives for others to punish as when the sum of this vector's sanctions is applied in one period.

More concretely, we will show below that – in an environment with one punishment stage – there are only three possible equilibria of the punishment game. First, if contributions do not give rise to “too much” inequity/inequality, no one will apply sanctions. If there is “too much” inequity (inequality), the inequity-averse (inequality-averse) enforcer will punish to equity (equality) if other players accept the arising inequality (inequity) of this outcome. If they do not, further sanctions are applied that lead to a total destruction of payoffs, i.e., both the inequity- and inequality-averse enforcer sanction such that a final payoff of zero materializes. In a nutshell, this

³⁸ The arguments presented below also hold for a proportional punishment technology.

³⁹ There is one notable exception. Punishment can only be applied in stage n if some punishment points were used in stage $n-1$. This does, however, not change the fact that one stage of punishment leads to the same incentives as punishment spread out over multiple stages.

is true since punishing to less than the equity (equality) benchmark cannot be an equilibrium outcome. *Unilaterally* deviating and erasing the remaining inequity (inequality) by punishing more would always be beneficial since it can be done at zero cost (as the punishment has already been paid). Additionally, in a situation in which both the high-return and the low-return enforcer have invested in sanctions, they always have a *unilateral* incentive to sanction more if there is remaining inequity or inequality (as this is associated with zero direct costs). Only when payoffs are zero is there no remaining inequality and inequity.

Does introducing a second (or multiple) punishment stages change the incentives for initial and retaliatory sanctions. First, if contributions do not provide incentives to punish due to “too much” inequity or inequality, nothing changes. Second, if there is e.g. “too much” inequity, the high-return enforcer still has an incentive to sanction to equity in the first punishment stage⁴⁰, especially if the low-return enforcer accepts these sanctions. Third, if the low-return enforcer does not accept sanctioning to equity, his or her potential counter-sanctioning in stage 2 could – in principle – provide the high-return enforcer with an incentive to punish more moderately initially and to accept some remaining inequity. Crucially, however, such a deterrent effect cannot be part of a subgame perfect equilibrium, as paying the punishment flat fee in one stage allows for costless punishment in all other stages and every player can punish in any stage: More moderate sanctions of the high-return enforcer in stage 1 (so that the low-return enforcer’s counter-sanctions are avoided) imply that some inequity remains in stage 2. Any equilibrium of the stage 2 subgame, however, must involve the high-return enforcer erasing this inequality, as s/he can do so at zero cost, effectively leading to the same outcome as in the one-stage punishment environment.⁴¹

Overall, while our framework with only one stage of punishment is unable to predict a sequence of retaliatory actions (over several punishment stages), we will characterize equilibria below in which one party’s sanctioning creates an incentive for another party to sanction as well, which in turn creates incentives to sanction even more for the first party, leading to a total destruction of payoffs. We refer to such a pattern as feuding. In an environment with multiple punishment stages, such a destruction of payoffs could always be achieved over multiple stages. In other words, the counter-sanctions that are immediately applied in an environment with one stage of punishment could be spread out over multiple stages of punishment in the other environment.

⁴⁰ Of course, the sanctions could also be divided and applied over multiple stages.

⁴¹ Erasing the remaining inequality is, of course, associated with potentially large indirect costs as it provides the low-return enforcer with an incentive to sanction as well. Nonetheless, not erasing remaining inequity is not credible as unilaterally sanctioning more is without direct cost.

Part (a).

(i) Note first, that if everyone contributes according to the schedule $\bar{c}$ no sanctions will be meted out. A player that is sufficiently inequality-averse, often called the (conditional cooperative) *enforcer* in the literature, does not suffer any disutility due to inequality and, thus, has no reason to sanction.

Let us now consider deviations from $\bar{c}$. We will assume that the deviator j is a low-return player.⁴² Player j deviates by contributing $\bar{c}_l - \Delta$. If all other players contribute $\bar{c}$ and the conditional enforcer sticks to the sanctioning strategy, then the deviator j gets the same monetary payoff as the enforcer i . We assume here that cost of punishment is a fixed fee and is normalized to 1, reflecting the cost of punishment in our experiment.⁴³ In this case, monetary payoffs are given by

$$\begin{aligned}\pi_i &= 20 - \bar{c}_h + m_h \left[\sum_{k=1}^n \bar{c}_k - \Delta \right] - 1 \\ \pi_j &= 20 - (\bar{c}_l - \Delta) + m_l \left[\sum_{k=1}^n \bar{c}_k - \Delta \right] - \Delta(1 + (m_h - m_l)) - 1 \\ &= 20 - \bar{c}_l + m_l \left[\sum_{k=1}^n \bar{c}_k \right] + \Delta - m_l \Delta - \Delta - m_h \Delta + m_l \Delta - 1 \\ &= 20 - \bar{c}_h + m_h \left[\sum_{k=1}^n \bar{c}_k \right] - m_h \Delta - 1 = 20 - \bar{c}_h + m_h \left[\sum_{k=1}^n \bar{c}_k - \Delta \right] - 1 = \pi_i\end{aligned}$$

For these calculations it has been assumed that in a symmetric (or *semi*-symmetric) equilibrium all high-return and all low-return players have to provide the same contribution $\bar{c}_h$ and $\bar{c}_l$ respectively. Since payoffs are identical, these contributions lead to the same payoff for both types (*) $20 - \bar{c}_h + m_h \sum_{i=1}^n \bar{c}_i = 20 - \bar{c}_l + m_h \sum_{i=1}^n \bar{c}_i$. Notably, the deviator's "initial" gain from deviating is $(1 - m_l)\Delta$. This gain is, however, overcompensated by the sanctions of $\Delta(1 + \hat{m}) + 1$. Overall, a net loss of $\Delta(1 + \hat{m}) + 1 - (1 - m_l)\Delta = 1 + \Delta m_h$, making the deviation unprofitable.

⁴² The case of a high benefit player is very similar to the homogenous case.

⁴³ The specific conditions presented in Proposition 2 are obviously shaped by the fact that we assume (and implement) a flat fee of punishment. Notably, a constant marginal punishment cost would still lead to qualitatively similar results in part a of the proposition as long as it is sufficiently cheap (a 1-1 punishment technology would not work) enough such that inequality-averse players can gain from punishing. In part b, the flat fee implies that players will only sanction as long as the level of inequality (or inequity) is high enough to recuperate this cost. Such a pattern could also be found with a concave punishment-cost function, in which punishment might only be rational as long as a sufficient amount of inequality (or inequity) is erased.

We now have to check whether sanctions are credible, whether the enforcer is really better off from sanctioning compared to not sanctioning. Thus $u_i(\bar{c}, p > 0) > u_i(\bar{c}, p = 0)$. Note first, that due to the deviation all high-return players lose m_h . As outlined before, the deviator gains $(1 - m_l)$. Finally, the other low-return players lose m_l . Thus, if the enforcer does not sanction, not only does the deviator have a higher payoff of $(1 + \hat{m})\Delta$, but all of the other low types will have a higher payoff of $\hat{m}\Delta (= (m_h - m_l)\Delta)$. For this reason, the following must hold:

$$\begin{aligned} u_i(\bar{c}, p > 0) &= 20 - \bar{c}_h + m_h \left[\sum \bar{c}_k - \Delta \right] - 1 \\ &\geq 20 - \bar{c}_h + m_h \left[\sum \bar{c}_k - \Delta \right] - \frac{\alpha}{n-1} \Delta [1 + \hat{m} + (n_l - 1)\hat{m}]. \end{aligned}$$

This inequality holds if $1 \leq \frac{\alpha}{n-1} \Delta [1 + n_l \hat{m}]$. Since the smallest possible deviation is one unit in the experiment, this inequality holds for any possible deviation, if $\alpha \geq \frac{n-1}{1+n_l \hat{m}}$. In addition, the enforcer cannot gain from deviating in the contribution stages as long as the same regularity condition with respect to β as in proposition 1 is fulfilled. It is easy to show that deviating by choosing $c_i > \bar{c}$ cannot be profitable for any player.

(ii) Note first that if all players contribute according to the schedule $\bar{c}$ no sanctions will be meted out. The enforcer does not suffer any disutility due to inequity and, thus, has no reason to sanction. Suppose that a low-return player j deviates by $\bar{c}_l - \Delta$. If all other players contribute $\bar{c}$ and the conditional enforcer sticks to the sanctioning strategy, then the deviator j gets $\frac{m_l}{m_h}$ times the payoff of enforcer i : Compared to a situation of no deviation, the enforcer first of all gets $m_h \Delta$ less payoff due to lower contributions. In addition, the enforcer loses 1 because s/he applies sanctions. To ensure that s/he still gets $\frac{m_h}{m_l}$ times the payoff of the deviator; the enforcer has to ensure that – compared to the situation without any deviation – the deviator's payoff is reduced by $\frac{m_h \Delta + 1}{m_h/m_l}$. Since the deviation as such already increases the deviator's payoff by $(1 - m_l)\Delta$, the enforcer has to sanction by $\Delta \left[(1 - m_l) + \frac{m_h}{m_h/m_l} \right] + \frac{1}{m_h/m_l} = \Delta + \frac{m_l}{m_h}$. These sanctions ensure that the deviation is not profitable for the deviator as $(1 - m_l)\Delta - \left(\Delta + \frac{m_l}{m_h} \right) = -m_l \Delta - \frac{m_l}{m_h} < 0$. In addition, the payoff of high-return and low-return players that do not deviate will be adjusted. While high-return players lose the same amount due to the deviation as the enforcer, their earnings have to be reduced by 1 (reflecting the punishment cost). In addition, the enforcer ensures that low types (that have not

deviated) will also lose $\frac{m_h \Delta + 1}{m_h/m_l}$ compared to the non-deviation payoff. As low types already lose Δm_l , these sanctions turn out to be $\Delta \left[\frac{m_h}{m_h/m_l} - m_l \right] + \frac{1}{m_h/m_l} = \frac{m_l}{m_h}$.

We now have to check whether the enforcer is really better off from sanctioning compared to not sanctioning. Thus $u_i(\bar{c}, p > 0) > u_i(\bar{c}, p = 0)$. If the enforcer does not sanction, his or her payoff is Δm_h lower than without any deviation. In addition, the deviator's payoff is $\Delta(1 - m_l)$ higher than before, implying a disutility of $\frac{\gamma_i}{n-1} \Delta \frac{1}{m_l} [m_h(1 - m_l) + m_l m_h] = \frac{\gamma_i}{n-1} \frac{m_h}{m_l} \Delta$. Notably, without sanctions, all the remaining high-return players will have a payoff that is exactly equal to the enforcer's payoff. In addition, no further disutility arises for the other low-return players that do not deviate. While the enforcer loses Δm_h , the other low types loose Δm_l and it holds that $-m_h m_l + m_l m_h = 0$. Thus, overall, we have

$$\begin{aligned} u_i(\bar{c}, p > 0) &= 20 - \bar{c}_h + m_h \left[\sum_{k=1}^n \bar{c}_k - \Delta \right] - 1 \\ &\geq 20 - \bar{c}_h + m_h \left[\sum_{k=1}^n \bar{c}_k - \Delta \right] - \frac{\gamma_i}{n-1} \Delta \left[\frac{m_h}{m_l} \right] = u_i(\bar{c}, p = 0). \end{aligned}$$

This inequality holds if $1 \leq \frac{\gamma_i}{n-1} \Delta \left[\frac{m_h}{m_l} \right]$. Since the smallest possible deviation is one unit in the experiment, the inequality holds for any possible deviation if $\gamma_i \geq \frac{n-1}{m_l}$. In addition, the conditional enforcer cannot gain from deviating in the contribution stages as long as the same regularity condition with respect to γ as in proposition 1 is fulfilled. Of course, it is easy to show that deviating by choosing $c_i > \bar{c}$ cannot be profitable for any player.

The case that a high-return player deviates is similar, but even a bit simpler. A high-type deviator gains $\Delta(1 - m_h)$ while the enforcer still loses Δm_h . Thus, taking the cost of punishment into account, sanctioning the deviator by $\Delta + 1$ will restore equity. In addition, it must hold that

$$u_i(c, p > 0) = 20 - \bar{c}_h + m_h \left[\sum_{k=1}^n \bar{c}_k - \Delta \right] - 1 \geq 20 - \bar{c}_h + m_h \left[\sum_{k=1}^n \bar{c}_k - \Delta \right] - \frac{\gamma_i}{n-1} \Delta [1]$$

This inequality holds if $1 \leq \frac{\gamma_i}{n-1} \Delta [1]$, implying a slightly stricter condition than for the low-type deviator: $\gamma_i \geq n - 1$.

Part (b).

We will first consider the case where all group members act according to the contribution schedule of i , i.e., high-return players contribute fully and low ones contribute c^{min} . For $c^{min} < c^{max}$, we show that no player has an *incentive to sanction*. The underlying idea is that if the high-return player's aversion to disadvantageous inequality γ_i is sufficiently low, the disutility arising from the contribution schedule from inequality is lower (or at most equal) than the cost of punishment. In other words, the γ_i determines the minimum level of contributions from low-return individuals the high-return enforcer is willing to accept, c^{min} , given all high-return players contribute fully. Similarly, if the low-return enforcer's aversion to disadvantageous inequality α_i is sufficiently low, the disutility arising from the contribution schedule from inequality is lower (or at most equal) than the cost of punishment. In other words, the α_i determines the maximum level of contributions low-return players are willing to contribute under the threat of punishment, c^{max} , given that all high-return players contribute fully.

To judge how low α_i has to be such that the low-return enforcer does not invest in sanctions and accepts the inequality associated with a particular contribution schedule, we have to describe the difference in payoffs between high and low-return players in terms of contributions. If high-return players contribute fully and low-return players contribute $\tilde{c}$, high types earn $m_h(n_h 20 + n_l \tilde{c})$ while low types earn $20 - \tilde{c} + m_l(n_h 20 + n_l \tilde{c})$. Thus, from the perspective of a low-return enforcer, (disadvantageous) inequality with n_h high types arises: $\frac{\alpha_i}{n-1} n_h (\hat{m}(n_h 20 + n_l \tilde{c}) - (20 - \tilde{c}))$. In addition, we assume that the low-return enforcer will only sanction when there is the minimal possible deviation of $1 - \hat{m}$, implying $\frac{\alpha_i}{n-1} [n_h (\hat{m}(n_h 20 + n_l \tilde{c}) - (20 - \tilde{c})) + 1 - \hat{m}]$. We first consider the extreme case of $\tilde{c} = 20$, i.e., everyone including low-return players contribute fully. Equating the utility cost with the normalized cost of punishment of 1, implies the following condition: $\alpha^{low} = \frac{n-1}{20n\hat{m}n_h+1-\hat{m}}$. Only if the low-return enforcer's aversion against inequality is sufficiently high, $\alpha_i > \alpha^{low}$, would s/he evaluate the inequality arising from an efficient contribution schedule as high enough so that sanctions would be considered beneficial. More generally, equating the disutility due to inequality with the punishment cost of 1 and solving for low-return players' contribution, provides the maximum level of own contributions the low-return enforcer (with a disutility parameter α_i) is willing to contribute (if forced to do so):

$$c^{max} = \left(\frac{n-1}{\alpha_i n_h} - \frac{1}{n_h} (1 - \hat{m}) - 20(\hat{m}n_h - 1) \right) \frac{1}{\hat{m}n_l + 1}. \quad (\text{B3})$$

As before, one can calculate the disutility a high-return enforcer incurs if high-return players contribute fully and low-return players contribute $\tilde{c}$. Given the payoff differences as calculated above, the disutility due to (disadvantageous) inequity is given by

$$\begin{aligned} \frac{\gamma_i}{n-1} \left[n_l \frac{1}{m_l} \left(m_h (20 - \tilde{c} + m_l (n_h 20 + n_l \tilde{c})) - m_l m_h (n_h 20 + n_l \tilde{c}) \right) + 1 \right] \\ = \frac{\gamma_i}{n-1} \left[n_l \left(\frac{m_h}{m_l} (20 - \tilde{c}) \right) + 1 \right], \end{aligned}$$

where the “+1” incorporates the feature that contributions in our experiment are discrete and we assume that the high-return enforcer will only sanction if the smallest possible deviation from his or her level of tolerance occurs. For the extreme case of equal earnings, this implies the following condition: $\gamma_i^{low} = \frac{n-1}{n_l} \left[\left(\frac{m_l}{m_h} \frac{1}{20 - \tilde{c}^{eq}} \right) + n_l \right]$, where $\tilde{c}^{eq} \left(= \frac{20 - n_h 20 \hat{m}}{1 + n_l \hat{m}} \right)$ is the low-return players’ contribution that ensures equal earnings. Thus, if the high-return player’s aversion against inequity is sufficiently small, $\delta_i < \delta^{low}$, s/he will even accept the equality rule. More generally, one can derive the minimum contribution that the high-return enforcer (with a disutility parameter γ_i) requires by equating his utility cost due to inequity to the cost of punishment and solving for his or her contribution:

$$c^{min} = \left(\frac{20m_h}{m_l} - \frac{n-1}{\gamma_i n_l} + \frac{1}{n_l} \right) \frac{m_l}{m_h}. \quad (\text{B4})$$

With these results in mind, let us consider what happens in case players deviate. For a better understanding let us first consider the extreme examples. First, let the low-return enforcer’s aversion against inequality be so weak that s/he is even willing to accept full contributions from everyone. If one player deviates from this schedule, e.g. a low-return player (or even the low-return enforcer), punishment is meted out by the high-return enforcer exactly as described in part a (ii) of the proposition. These kind of sanctions ensure that equity is fully restored although the level of earnings is slightly reduced for everyone (accounting for the cost of punishment). For this reason, however, the overall level of inequality after sanctions are applied cannot be higher than that resulting from full contributions. Therefore, even if the deviator is the low-return enforcer, s/he has no incentive to deviate and engage in sanctioning (as s/he would have been willing to accept the original earnings distribution) and is worse off.

Second, let us consider the case where contributions are such that equal earnings materialize and the high-return enforcer is willing to accept the outcome while the low-return enforcer strictly insists on equal payoffs. Deviations from this contribution schedule will be sanctioned in a very similar way than in part a (i) of the proposition. The only difference is that the enforcer this time is

a low-return player, leading to smaller differences: The enforcer i chooses $p_{ij} = \Delta (1 - \hat{m}) + 1$ if the deviator is a high type and $p_{ij} = \Delta + 1$ if the deviator is a low type as well as $p_{ik} = 1$ for $k \neq j$ if k is low type and $p_{ik} = 1 - \Delta \hat{m}$ if k is a high type. In addition, the α -condition is slightly different compared to Proposition 2 due to the fact that the enforcer is a low type: If a high type deviates by contributing Δ less, the deviator will earn $(1 - m_h)\Delta$ more while the enforcer will have m_l less, implying an increased disutility of disadvantageous inequality of $\frac{\alpha_i}{n-1}(1 - \hat{m})$ and, thus, a condition of $\alpha_l^{high} \frac{n-1}{1-\hat{m}}$. Otherwise, however, the logic is the same.

Let us now consider the general case that the enforcers' α_i and γ_i are such that $c^{min} \leq c^{max}$. A contribution schedule in which high-return individuals contribute fully and low-return individuals contribute c^{min} will not lead to any sanctions as the enforcers' aversions to inequality and inequity are sufficiently low to tolerate the associated level of inequity and inequality.

We will show next that deviations from this schedule are also not profitable and the associated sanctions are credible. If a deviation of a low-return player occurs, one of the following equilibria of the punishment game will be played. First, the high-return enforcer will sanction such that from his or her point there is no inequity any more. This sanctioning is similar to the sanctioning in part a (ii) of this proposition but has to be modified slightly to take into account that the benchmark without any deviation is already characterized by some inequity. In case a deviation Δ occurs, the (high-return) enforcer's payoff after punishment cost is: $\pi_h^{enf} = m_h(n_h 20 + n_l c_{min}) - m_h \Delta - 1$. To erase inequity, the enforcer has to ensure that other high-return players earn the same payoff π_h^{enf} while low-return players must earn $\pi_h^{enf} / \frac{m_h}{m_l}$. Since the deviators payoff (after the contribution stage) is $20 - c_{min} + m_l(n_h 20 + n_l c_{min}) + (1 - m_l) \Delta$, the enforcer can ensure his or her objective by choosing $p_{ij} = 20 - c_{min} + \Delta + \frac{m_l}{m_h}$ for the deviator as well as $p_{ik} = 1$ for other high types and $p_{ik} = 20 - c_{min} + \frac{m_l}{m_h}$ for other low types. Importantly, this sanctioning schedule can only be part of an equilibrium strategy if the low-return deviator is selfish or if the deviator is equality-minded, but his or her aversion to inequality is sufficiently small to accept the arising inequality. In particular, this inequality has to be smaller than the one associated with low-return individuals playing c^{max} . Since the low-return player is sanctioned to $\pi_h^{enf} / \frac{m_h}{m_l}$, and the high-return player earns π_h^{enf} , the resulting inequality is $\pi_h^{enf} (1 - \frac{m_l}{m_h})$. Thus, our condition is

$$\begin{aligned}
& (m_h(n_h 20 + n_l c^{min}) - m_h \Delta - 1) \left(1 - \frac{m_l}{m_h}\right) \\
& \leq n_h (\widehat{m}(n_h 20 + n_l c^{max}) - (20 - c^{max})) .
\end{aligned} \tag{B5}$$

Of course, the higher c^{max} is the higher is the inequality the low-return enforcer is willing to accept and the more space there is for the high-return player's sanctioning. If the condition is, however, not fulfilled, the low-return enforcer has an incentive to deviate in the punishment stage and invest in punishment points. Notably, the high-return enforcer has still no incentive to punish less. First, for a lower quantity of sanctions, it might happen that the reduction in inequity is not sufficient to recover the punishment cost. But even if this is not the case, not reducing *all* potential inequity (to accommodate the low-return player) cannot be an equilibrium strategy as there is a flat fee of punishment. In a situation in which some inequity remains, the high-return player would always have a *unilateral* incentive to deviate (even though more sanctioning may invoke counter-punishment from the low-return enforcer).

If the condition is not fulfilled, the following equilibrium emerges: Both the high-return and the low-return enforcer invest in sanctions. While the high-return enforcer sanctions such that low-return players have a payoff of zero, the low-return enforcer ensures that high-return players have a payoff of zero. A payoff of zero ensures that there is no inequality or inequity. While the implications of this strategy with respect to payoffs are obviously severe, these strategies are still credible: In case the low-return enforcer acts as described, it is a best response for the high-return enforcer to destroy the low-return player's payoff (to reduce inequity) and vice versa. In case the condition is not met, it is actually the unique equilibrium. First, not investing in sanctions, by both the low-return and the high-return enforcer, cannot be equilibrium behavior because the high-return enforcer would have a unilateral incentive to deviate in the punishment stage and to invest to punish. Second, while one enforcer tries to ensure equality, the other tries to maintain equity. These incentives only vanish if everyone earns zero. As a response, players should avoid contributions that would lead to this outcome.

After considering deviations from low-return players, we will now come to deviations from high-return players. Of course, the high-return enforcer can sanction deviations by other high-return players, for example selfish ones, in a similar manner to that described earlier. It is, however, important to note that a simple condition on γ [as in part (a) of the proposition] cannot ensure that the high-return enforcer him- or herself contributes fully. As long as low-return players do not contribute fully, contributing fully is inevitably associated with disadvantageous inequity for high-return players. If $c^{max} = c^{min}$, the low-return enforcer could sanction the high-return deviator in a similar fashion to that described earlier. But even if the low-return enforcer is willing to accept

quite a lot more inequality than the high-return enforcer is willing to accept inequity ($c_{min} \ll c_{max}$), an equilibrium in which high-return individuals contribute fully and low-return individuals c_{min} can still be maintained as long as there is a second high-return enforcer (with the same c^{min}). In this case, any deviation by one of the high-return enforcers increases the overall inequity so that the other high-return enforcer has an incentive to sanction according to the above description.⁴⁴ While there are other possible modelling approaches,⁴⁵ we believe that a second high-return enforcer is plausible as our design biases in favor of efficiency – due to the real-effort task – making it more likely that high-return players are particularly willing to sanction to achieve their objectives.⁴⁶

Are there any other (sanction-free) symmetric equilibria on the Pareto-frontier for the case that $c^{min} \leq c^{max}$? Note first that if low-return players contribute less than c^{min} , sanctions will occur as the level of inequity is too high for the high-return enforcer. If they contribute more, selfish low-return players have an incentive to lower their contributions. The only exception is in case that (all) low-return players contribute c^{max} . If selfish low-return players deviate downwards, this increases inequality for the low-return enforcer. If low-return players contribute c^{max} , the maximum acceptable level of inequality has already been reached, implying that the low-return enforcer would punish the deviation. Of course, even in this situation the only low-return enforcer may still have an incentive to deviate downwards him- or herself (and a selfish player will not punish him for that). Generally, let us consider the case that all high-return players contribute fully and all low-return player contribute $\tilde{c} > c^{min}$. If the low-return enforcer deviates down by Δ , s/he will directly gain $(1 - m_l) \Delta$ in payoffs. In addition, disadvantageous inequality (with respect to n_h high-return

⁴⁴ If such a second enforcer exists, the punishment strategies described above for the case of a low-return deviator have to be adjusted such that e.g. both enforcers equally share in the punishment of the other players.

⁴⁵ One could also introduce a second low-return enforcer. In this case $c_h = 20, c_l = c^{max}$ can, in principle, be an equilibrium outcome if both low-return enforcers are able to discipline each other in contributing enough. While the qualitative message would still be very similar – especially with respect to the comparative statics between conditions/proposition – we choose to assume a second high-return enforcer instead of a low-return enforcer for the two reasons outlined in the main text. In addition, a second low-return enforcer can be somewhat less effective than a second high-return enforcer: If there are two enforcers of the latter type and one of them deviates by contributing Δ less, this increases inequality for the other enforcer by Δ , providing a reason to punish. If there are two low-return enforcers and one of them deviates by contributing Δ less, there will still be a similar increase of Δ in inequality. Unlike before, however, lower contributions from a low-return player decreases inequality with respect to all high-return players by $n_h(m_h - m_l)\Delta = n_h\hat{m}\Delta$ (as high-return individuals benefit disproportionately from any contributions). Thus, especially in cases with many high-return players, even two low-return enforcers would not be able to punish a deviation from another enforcer. Notably, even with two low-return enforcers, intermediate outcomes ($c_h = 20, c^{min} < c_l < c^{max}$) can still not be justified as (semi-symmetric) equilibrium play: For an intermediate low-return contribution, low-return players have an incentive to contribute less, as neither high-return nor low-return enforcers have an incentive to punish if the deviation is not too large.

⁴⁶ In addition, empirically, it may be more likely that high-return players manage to contribute fully even in the absence of a threat of punishment from a low-return player (than vice versa) as the high-return players' joint surplus from the public account is larger than one in our experiment.

players) is reduced by $\alpha_i \frac{n_h}{n-1} (1 + \hat{m}) \Delta$, but advantageous inequality (with respect $n_l - 1$ fellow low-return players) increases by $\beta_i \frac{n_l-1}{n-1} \Delta$. Thus if $\left[(1 - m_l) + \alpha_i \frac{n_h}{n-1} (1 + \hat{m}) - \beta_i \frac{n_l-1}{n-1} \right] \Delta < 0$, low-return player would have no incentive to deviate, resembling similar conditions in Proposition 1. Of course, since $\alpha_i \geq \beta_i$, and we have an equal number of high and low-return players in our experiment, this can never happen under our experimental parametrization.

Note additionally, that we focus on symmetric equilibria. Of course, other equilibria exist in which the overall level of inequity that is acceptable is distributed among players in a different way. We believe, however, that the symmetric equilibria are more focal. Finally, there can be symmetric equilibria that are not on the Pareto-frontier, i.e. in which high return players do not contribute fully. We will discuss these equilibria in more detail below (for the case that aversions to inequity and inequality are not compatible). Of course, in the presence of equilibria on the frontier, these inferior equilibria are not very attractive.

Finally, if $c^{max} < c^{min}$, there cannot be an equilibrium without sanctions on the Pareto-frontier. Any contribution of low-return players below c^{min} would imply sanctions from the high-return enforcer, while any contribution above c^{max} would not be acceptable to the low-return enforcer. In the following, we will establish the circumstances under which Pareto-inferior contributions can be part of an equilibrium without sanctions. This includes the case of zero contributions. If these circumstances are not met, sanctions are inevitable. In the next step, we will further analyse under which conditions a destruction of payoffs and, thus, a feud is inevitable.

To establish when equilibria (without sanctions) exist that are not on the Pareto-frontier, we have to consider contribution schedules that have the same level of inequality or inequity as those on the frontier and analyze whether these lines of equal inequity and equal inequality intersect. If they do intersect, it means that there are contributions schedules (below the frontier) for which both the levels of inequity and inequality are acceptable. Let us first determine all (symmetric) contribution schedules that result in the same amount of inequality $\hat{I}$ that we observe when high-return players contribute fully and low-return player contribute c^{max} . Between one high-return and one low-return player, the following degree of inequality emerges: $\hat{I} = \pi_h(20, c^{max}) - \pi_l(20, c^{max}) = m_h(20n_h + n_l c^{max}) - m_l(20n_h + n_l c^{max}) - (20 - c^{max}) = \hat{m}(20n_h + n_l c^{max}) - (20 - c^{max})$. In the special case of $c_l = c^{eq}$, $\hat{I} = 0$. Let us consider the case where all high-return players contribute c_h and all low-return players contribute c_l and the same amount of inequality arises, $\pi_h - \pi_l = \hat{I}$. This implies that $m_h(n_h c_h + n_l c_l) + (20 - c_h) - m_l(n_h c_h + n_l c_l) - (20 - c_l) = \hat{I}$. Solving for c_l :

$$c_l = \frac{\hat{I}}{1 + \hat{m}n_l} + \frac{1 - \hat{m}n_h}{1 + \hat{m}n_l} c_h = \frac{\hat{m}(20n_h + n_l c^{max}) - (20 - c^{max})}{1 + \hat{m}n_l} + \frac{1 - \hat{m}n_h}{1 + \hat{m}n_l} c_h. \quad (\text{B6})$$

Thus, there exists a set of (symmetric) contributions that all lead to inequality of $\hat{I}$. For the extreme case that low-return players are only willing to accept equal payoffs ($c^{max} = c^{eq}$), $\hat{I}$ is zero and we simply have $c_l = \frac{1 - \hat{m}n_h}{1 + \hat{m}n_l} c_h$. For the parameters in our experiment, the slope is 0.25, implying that whenever high-return player contribute four times as much as low-return one, equal payoffs arise.

In a second step, we will determine all (symmetric) contribution schedules with the same amount of inequity $\bar{I}$ that we observe when high-return players contribute fully and low-return players contribute c^{min} . Between one high-return and one low-return player, the following degree of (weighted) inequity emerges: $\bar{I} = m_h \pi_l (20, c^{min}) - m_l \pi_h (20, c^{min}) = m_h [m_l (20n_h + n_l c^{min}) - (20 - c^{min})] - m_l [m_h (20n_h + n_l c^{min})] = m_h (20 - c^{min})$. Of course, if high-return player require efficient or full contribution ($c^{min} = 20$), there will not be any inequity, and $\bar{I} = 0$. Let us consider the case when all high-return players contribute c_h and all low-return players contribute c_l and the same amount of inequity arises, $m_h [m_l (n_h c_h + n_l c_l) + (20 - c_l)] - m_l [m_h (n_h c_h + n_l c_l) + (20 - c_h)] = \bar{I}$. Solving for c_l shows the relationship between low- and high-return players' contributions:

$$c_l = 20 - \frac{m_l}{m_h} 20 - \frac{\bar{I}}{m_h} + \frac{m_l}{m_h} c_h = c^{min} - \frac{m_l}{m_h} 20 + \frac{m_l}{m_h} c_h. \quad (\text{B7})$$

For the extreme case where high-return players are only willing to accept full contributions ($c^{min} = 20$), we have $c_l = 20 - \frac{m_l}{m_h} 20 + \frac{m_l}{m_h} c_h$. For the parameters in our experiment, the slope is 0.5 and the intercept is 10, implying that a situation in which high-return players contribute nothing but low-return player contribute 10 is as acceptable as everyone contributing fully. This illustrates that inequity-averse players will not sanction as long as they have adequately higher payoffs than low-return players independently of the level of earnings. Of course, their utility is still higher when everyone contributes fully.

We have to consider two cases: If the slope of the line of equal equity is smaller than the line of equal equality, $\frac{m_l}{m_h} < \frac{1 - \hat{m}n_h}{1 + \hat{m}n_l}$, it is clear that $c^{max} < c^{min}$ is already sufficient to rule out any equilibria without sanctions even if we look below the Pareto-frontier. The second case, $\frac{m_l}{m_h} > \frac{1 - \hat{m}n_h}{1 + \hat{m}n_l}$, is, however, the one of our experimental parameters. Here, even though aversions against inequity and inequality are incompatible on the frontier, the two lines can still intersect and players

could agree on a contribution schedules in which high-return players do not contribute 20. Equating the intercept of both lines provides a condition for how much higher c^{min} has to be compared to c^{max} , so that the two lines do not intersect for (positive contributions). We get:

$$c^{min} > \frac{\hat{m}(20n_h + n_l c^{max}) - (20 - c^{max})}{1 + \hat{m}n_l} + \frac{m_l}{m_h} 20 (> c^{max}). \quad (\text{B8})$$

For the parameters of our experiment, the formula implies that the condition reduces to : $c^{min} > c^{max} + 5$. Generally speaking, as long as tolerance levels lead to c^{min} 's and c^{max} 's that are sufficiently apart from one another, groups cannot settle on equilibria (without sanctions) that are below the frontier. Notably, both the equality rule ($c_h = 20, c_l = c^{eq}$) and zero contributions ($c_h = c_l = 0$) feature the same level of inequality (zero). This implies that inserting $c^{max} = c^{eq}$ in the above condition provides a condition for when high-return players are unwilling to accept zero contributions (without sanctions). Their aversion to inequity has to be sufficiently high $c^{min} > \frac{m_l}{m_h} 20$. Whenever the conditions is met, zero contributions (without sanctions) cannot be rationalized as an equilibrium. For the parameters of our experiment, this implies $c^{min} > 10$.

So far, we have established conditions under which any equilibrium has to involve sanctions. Intuitively, if the tolerance levels of high-return and low-return players are too far apart, $c^{max} \ll c^{min}$, any contribution schedule will result in subsequent sanctioning as either the associated level of inequality or inequity is too high (and this even holds for zero contributions). Of course, whether e.g. a high-return player's sanctions creates an incentive for low-return player to sanction as well depends on how much inequality these sanctions create and how much tolerance the low-return enforcer has. For intermediate values of inequity and inequality aversion, both contribution schedules in which some punishment is meted out and in which payoffs are completely erased can potentially be rationalized in equilibrium. Here, the former equilibria are obviously more attractive than the latter ones. In the following, we will not characterize all these equilibria but focus on conditions under which any contribution schedule will lead to a complete destruction of payoffs. Intuitively, this will be the case when the low-return enforcer is strongly *inequity* averse and the high-return enforcer is strongly *inequality* averse. As an example, consider the case that the high-return enforcer is willing to sanction any deviation from equity (or $c^{min} = 20$) while low-return players are willing to sanction any deviation from equality (or $c^{max} = c^{eq}$). It is clear that any contribution schedule will create incentives to sanction following the conditions established above. These sanctions are, however, either aimed at restoring equality or equity and, thus, will always create either inequity or inequality, generating an incentive for even more sanctioning from the

other side. For this reason, payments will be eroded unless everyone has zero earnings so that equality and equity coincide.

Let us establish more precise conditions. The case of zero contributions shows that aversions against inequality and inequity can even be a bit less strong than the extrema. If all players contribute nothing, a high-return enforcer will sanction to equity as long as s/he is moderately inequity averse as indicated by the conditions outlined above (for our experimental parameters if $c^{min} > 10$). Will this create an incentive for low-return players to invest in punishment points as well? As long as the low-return players are not too lenient with accepting inequality (for our experimental parameters if $c^{max} < 12$), they will indeed invest in sanctioning, implying a total erosion of payoffs. Crucially, however, this equilibrium of the punishment game creates incentives to deviate with respect to initial contributions. Contributing zero with a complete destruction of players' payoffs will only be an equilibrium when players' aversion against inequality and inequity are strong enough to make any such deviation unprofitable. We consider deviations of a low-return player first: To avoid a destruction of payoffs, a low-return player has to contribute and sanction such that everyone gets the same payoff but the lowest possible inequity is inflicted on high-return players.⁴⁷ This is achieved by ensuring a very low payoff (for oneself) and everyone else. In other words, a low-return player could deviate from zero contributions, contribute fully and afterwards sanction to ensure that all players have the same payoff. This player will earn $20 \cdot m_l$ or $20 \cdot m_l - 1$ after applied sanctions. Since high-return players will earn the same amount, their disutility due to inequity aversion is:

$$\gamma \frac{n_l}{n-1} \frac{1}{m_l} [m_h(20 \cdot m_l - 1) - m_l(20 \cdot m_l - 1)] = \gamma \frac{n_l}{n-1} \left[1 + 20 \hat{m} - \frac{m_h}{m_l} \right].$$

If this disutility due to inequity is larger than 1, the high-return enforcer will sanction even this deviation. Crucially, as has been shown earlier, if high-return players contribute fully and low return players contribute c^{min} , the disutility due to inequity of $\gamma \frac{n_l}{n-1} \left[\frac{m_h}{m_l} (20 - c^{min}) \right]$ arises. Equating these two disutilities provides a condition for how inequity averse high-return players have to be to still have an incentive to sanction, i.e. $c_{feud}^{min} > 20 - \frac{m_l}{m_h} 20(m_h - m_l) - \frac{m_l}{m_h} + 1$. As long as $m_h > m_l$, c_{feud}^{min} will be smaller than 20. For the parameters of our experiment, this results in $c^{min} > 17.5$.

⁴⁷ As outlined above, sanctioning moderately and leaving some inequality (or inequity) on the table (so as not to provoke "counter-sanctions") cannot be part of an equilibrium sanctioning strategy as sanctioning more would always be a unilaterally beneficial.

In a second step, we will consider a deviation of a high-return player. How can s/he deviate (to avoid destruction of payoffs), reduce inequity and at the same time inflict the minimal possible inequality on low-return players? S/he acts in a similar way as the low-return deviator. Providing full contributions when everybody else contributes zero, leads to minimal possible earnings for the deviator of $20 \cdot m_h$ after the contribution stage. After the contribution stage, the deviator pays the flat punishment fee and applies sanctions to reach equity, leading to payoffs of $20 \cdot m_h - 1$ for high-return individuals and $20 \cdot m_l - m_l/m_h$ for low-return individuals, implying that the latter players feel a disutility of $\alpha_i \frac{n_h}{n-1} \left[20\hat{m} - 1 + \frac{m_l}{m_h} \right]$. As indicated before, contributing c^{max} leads to disutility due to inequality of $\alpha_i \frac{n_h}{n-1} [\hat{m}(20n_h + c^{max}n_l) - (20 - c^{max})]$. Equating these two disutilities provides a condition on how inequality averse low-return players have to be to still sanction the deviations, $c_{feud}^{max} < \frac{20(\hat{m}+1-n_h\hat{m})-1+m_l/m_h}{1+n_l\hat{m}}$. For the parameters of our experiment, this results is $c_{feud}^{max} < 7.18$.

Summarizing, if both high-return and low-return players are sufficiently inequity/inequality averse, there is an equilibrium in which zero contributions and a complete destruction of payoffs occur. Notably, any other equilibrium that might involve positive contributions will also lead to a destruction of payoffs if players have these strong preferences. The deviations considered above are particular in the sense that they lead to the lowest possible (post-contributions) payoffs for both high and low-return players and thus to the minimal possible infliction of inequity and inequality, respectively. If anything, deviations from equilibria involving positive contributions from other players will lead to a higher infliction of inequity or inequality. This is simply because positive contributions from other players increase the payoff of the punisher, implying that his or her deviation cannot reduce his or her own payoff level by too much. Thus, his or her subsequent punishment will inflict a higher level of inequity (or inequality). Overall, if one party is willing to accept almost no inequity and another party is willing to accept almost no inequality, the destruction of payoffs is inevitable. ■

The first part of proposition 2 shows that with (the mere threat) of punishment, selfish players can be disciplined to cooperate. In addition, the second part of Proposition 2 indicates that when communication and punishment act in tandem, high levels of cooperation can be sustained even in the presence of ex-post normative conflict, under some conditions. In particular, if the relative distastes for inequity and inequality are not too strong ($c^{min} \leq c^{max}$) a compromise between group members is possible. This also implies that even efficiency can be achieved if there are no strongly inequality averse individuals in the group. When the relative distaste for inequity and inequality is, however, too high ($c^{min} > c^{max}$), leaving no space for compromise on the Pareto-frontier, groups

have to settle for Pareto-inferior solutions in which high-return players do not contribute fully – potentially zero contributions – in case the relative distastes are not too far apart ($c^{\min} < c^{\max} + 5$ – for the parameter of our experiment). If preferences are too far apart ($c^{\min} > c^{\max} + 5$), groups face sanctions. Finally, if both sides are strongly averse to inequity and inequality – (for example $c^{\min} = 20, c^{\max} = 5$), a total destruction of payoffs – or in other words feuds – are inevitable.⁴⁸ Overall, as the requirements for the latter cases are fairly high, we would still expect groups to be better off overall than under communication alone.

B.4 Punishment without communication

In our analysis so far, we have seen that ex-post normative conflict might not always be resolvable, but at least, if punishment and communication are present jointly and preferences are not too extreme, there is a chance that groups might find a compromise or even reach efficiency. For this analysis, we have assumed that players have complete information about other players' preferences. However, in the absence of the ability to communicate, we presume that such complete information is not present. To illustrate the importance of ex-ante normative conflict, namely that conflict may arise simply due to uncertainty about the degree of other players' preferences or incomplete information, we will now model our experimental environment as a game of imperfect information in which nature decides on each player's degree of inequity/inequality. Notably, without communication, another problem arises, namely that players may find it more difficult to coordinate on Pareto-superior outcomes. While it is difficult to disentangle the two problems, appendix D provides a discussion regarding why we believe that (ex-ante) normative conflict poses the more serious problem.

Ex-post normative conflict. As in the main text, we will focus on a situation in which conflicting normative views coexist, i.e., ex-ante normative conflict emerges in the presence of ex-post normative conflict (but we will come back to the idea that ex-ante normative conflict may emerge even in the absence of ex-post conflict at the end of this section). As in the case of complete information, we assume that high-return players are *equity*-minded and low-return players are *equality*-minded. Players, however, do not know the exact degree of both aversions on the part of other players. Our point is a simple one: If (some) low-return players are unsure about the degree of high-return players' inequity aversion and underestimate it, they might contribute too little. If

⁴⁸ We finally note that the levels of α and γ that lead to punishment depend on the cost of punishment that we normalized to 1, according to the cost in our experiment. Of course, these levels might be higher insofar as sanctioning others may not only involve a monetary cost, but could also involve a psychological cost from which we abstract.

the beliefs of (some) high-return players still support full contribution, the insufficient contributions of (some) low-return players might even lead to feuds.

We formalize these ideas in a game of imperfect information, in which nature moves first and determines the exact degree of players' inequity and inequality aversion. In recognition of the many possible different equilibria in such a setting, we will choose the simplest possible setting to make our point. Thus, our analysis is less general than in the previous setting, but more closely aligned to what happens in our experiment, as in the experiment we have two high-return and two low-return players. For simplicity, we will model the situation with only two decision makers, one low and high-return player. The other two players exactly replicate what their respective same-return player does. This is done for tractability but also to make the underlying argument most transparent: We would like to analyze how low-return players' belief about the frequency of different high-return types affects their optimal contribution and this form of analysis puts the focus on Pareto-optimal behaviour.⁴⁹ Our aim is to show that plausible equilibria exist in which incomplete information can lead to suboptimal results (compared to an environment with complete information), as indicated earlier.⁵⁰

We will consider the following general environment. Nature assigns three potential degrees of *inequity* aversion to high-return player (*strong*, *medium* or *weak*) and two degree of *inequality* aversion to low-return players (*medium* or *weak*).⁵¹ Notably, the weaker the degree of *inequity* aversion of high-return players, the higher is their tolerance towards lower contributions from low-return players, i.e. the lower is the minimum contribution level, c^{min} that they require low-return

⁴⁹ With only two decision makers, the low-return (high-return) player's best response will only depend on the high-return (low-return) player's action and not how the other low-return (high-return) player behaves. Our focus will be to analyze how low-return players' beliefs about the proportion of high-return types determines their optimal behavior. With only two decision makers, we avoid equilibria in which low-return players only get stuck as the other low-return player contributes too little or too much even though a Pareto-superior solution would be attainable. For simplicity, we also assume that there are only four players overall (although this assumption could be easily relaxed).

⁵⁰ Notably, we also assume that – as in the experiment and in the previous analysis – the joint surplus of high-return players is higher than 1. This provides those players with an incentive to contribute fully (given that low-return players fulfil the contribution requirement c^{min}). For this reason, we assume – more precisely – that the joint return is sufficiently high to justify a ($c_h = 20$, $c_l = c^{min}$) equilibrium outcome. If (both) high-return players deviate by contributing Δ less, it would lead to a gain/loss in payoffs of $(1 - 2 \cdot m_h) < 0$. Crucially, inequity in such an outcome is characterized by $\gamma \frac{2}{3} [\frac{m_h}{m_l} (20 - c^{min})]$. A deviation of Δ reduces this inequity by $\gamma \frac{2}{3} [\frac{m_l}{m_l} \Delta]$. Thus, as long as the loss in payoff is higher than the gain in reduced disutility, high-return players have an incentive to contribute fully, $m_h > 1/3 \cdot \gamma + 1/2$. Empirically, in most groups high-return players manage to contribute fully, at least when punishment is available. Crucially, high-return players do not have an incentive to contribute fully under all circumstances. In case, low-return players e.g. do not fulfil the contribution requirement, punishment still has to be meted out as there is too much inequity, potentially making lower contributions beneficial.

⁵¹ Adding a strongly inequality-averse (low-return) player does not change the results qualitatively, but increases complexity without much further benefit.

players to contribute (while they contribute fully themselves – see Proposition 2). In contrast, the higher the degree of *inequality* aversion of low-return players, the lower is their willingness to contribute c^{max} (under the threat of punishment).

Simple setting. Before analysing this general environment, we want to make the simple point that incomplete information or uncertainty about the degree of aversions can lead to insufficient contribution in a simple environment. Afterwards, we will show how these insufficient contributions can even lead to feuds in the more general case. In our simple environment, only high-return players with *inequity* aversion of medium or weak degree exist. With probability p , high-return players are equity-minded to a weak degree, and with probability $1 - p$, they are equity-minded to medium degree. Low return players are always *equality*-minded (to a medium degree). For this reason, we will only talk about low-return or equality-minded players in the following discussion. More specifically we assume that aversions are ordered in the following way:

$$c_{m(edium)}^{max} = c_{m(edium)}^{min} > c_{w(eak)}^{min}$$

Focusing on equilibria on the Pareto-frontier, these (in-)equalities imply the following in a game of complete information: (low-return) individuals will *just* meet the high-return individuals respective contribution requirement, i.e. they will contribute c_m^{min} , or c_w^{min} if facing medium or weakly equity-minded high-return players, respectively (while those high types contribute fully).⁵²

In a game of incomplete information, high-return players know their own degree of inequality-aversion but low-return players only know the probability distribution of the degree of inequity-aversion. In other words, in choosing their contribution level, they have to maximize the expected payoff of the possible cases (i.e. we consider a Bayesian Nash equilibrium). In such a situation, low-return players may – under specific conditions – decide to fulfil only the more lenient contribution requirement. This can lead to situations (realizations) in which subjects' preferences would allow for a compromise contributions level, i.e. the levels of inequality and inequity aversion are compatible, or even efficiency, but low-return players nonetheless contribute too little, and in turn punishment can occur (although no feuds occur in our simple environment).

Proposition 3 (a): [Ex-Ante and Ex-Post Normative Conflict]

If the probability p that high-return players are weakly equity-minded is sufficiently high, i.e., if

⁵² Note, that we assume for simplicity that medium aversion against inequality and inequity exactly match ($c_m^{max} = c_m^{min}$). Of course, $c_m^{max} > c_m^{min}$ would lead to the same results as high-return players have an incentive to contribute fully in case low-return player fulfill the contribution requirement, as discussed in footnote 20. Compare to the previous footnote to see why high-return players always have an incentive to contribute fully (in case low-return players fulfill the contribution requirement).

$$p > u(20, c_m^{min}, p = 0) - u(b(c_w^{min}), c_w^{min}, p \geq 0) / u(20, c_w^{min}, p = 0) - u(b(c_w^{min}), c_w^{min}, p \geq 0),$$

fulfilling only the low requirement and contributing $\tilde{c} = c_w^{min} (< c_m^{min})$ can be rationalized as equilibrium play for low-return players. Whenever they meet high-return players that are only weakly inequity-averse, cooperation occurs ($c_{high} = 20, c_{low} = c_w^{min}$). Whenever they meet strongly equity-minded players, sanctions are meted out and/or contributions remain below the Pareto-frontier. Notably, contributing c_m^{min} can either not be rationalized as equilibrium play or, at least, leads to lower payoffs for low-return players.

Proof.

Note first, that both types of high-return players have an incentive to contribute fully in case inequality-averse players fulfill (or overfulfill) their respective contribution requirement. But what will the latter players do? Focusing on equilibria on the Pareto-frontier, they have an incentive to at least satisfy the more lenient contribution requirement and contribute c_w^{min} (in case weakly inequity-averse players contribute fully). In addition, if there are sufficiently many medium inequity-averse players, they have an additional incentive to also fulfill the more stringent requirement and contribute c_m^{min} .

Let $u_{l(ow)}^{m(edium)}(c_h^m, c_l, p)$ be the utility of low-return players, which depends on low-return and high-return players' contributions⁵³ as well as the punishment vector. By contributing c_m^{min} , players are able to ensure a (punishment-free) positive payoff, $u_l^m(20, c_m^{min}, p = 0)$, when meeting both medium and weakly inequity-averse high-return players. Contributing c_w^{min} will, however, only ensure a positive payoff that does not involve punishment for sure, $u_l^m(20, c_w^{min}, p = 0)$ for weakly inequity-averse players. The latter payoff is, of course, higher as the involved contribution of the low-return players is lower, $c_w^{min} < c_m^{min}$. When a more strongly equity-minded player is met, a utility of $u_l^m(b_h^m(c_w^{min}), c_w^{min}, p \geq 0)$ is gained, where $b_h^m(c_w^{min})$ is the best response of medium inequity-averse players towards a low-return players' insufficient contribution of c_w^{min} . As we will show below, those players have too main approaches to react to insufficient contributions of low-return players. Either they contribute fully and apply sanctions to reach equity afterward, or they provide less than full contributions and settle on a Pareto inferior solution. In both cases, the low-return players' utility will be below both $u_l^m(20, c_w^{min}, p = 0)$ and $u_l^m(20, c_m^{min}, p = 0)$.⁵⁴ Thus, if the

⁵³ Here, we omit contributions of weakly averse high-return players as this player type always has an incentive to contribute fully (as long as its contribution requirement is fulfilled).

⁵⁴ While the first inequality is trivial, the second holds for the following reason: High-return players earn less when low-return players only contribute c_w^{min} instead of c_m^{min} . Nonetheless, in finding their best-response $b_h^m(c_w^{min})$, high-return players will "exploit" low-return players' tolerance to inequality in a similar way as when the latter contribute c_m^{min} (see also below), implying that also low-return player will earn less.

number of weakly inequity-averse player is sufficiently high, contributing c_w^{min} leads to greater expected payoff, $p u_l^m(20, c_w^{min}, p = 0) + (1 - p)u_l^m(b(c_w^{min}), c_w^{min}, p \geq 0)$, than contributing c_m^{min} , $u_l^m(20, c_m^{min}, p = 0)$. This implies the condition:

$$p > \frac{u_l^m(20, c_m^{min}, p = 0) - u_l^m(b_h^m(c_w^{min}), c_w^{min}, p \geq 0)}{u_l^m(20, c_w^{min}, p = 0) - u_l^m(b_h^m(c_w^{min}), c_w^{min}, p \geq 0)}. \quad (\text{B9})$$

What will high-return players that are *inequity*-averse to a medium degree do? If low-return players decide to fulfil the requirement and contribute c_m^{min} , the optimal response is to contribute fully and there is no reason to apply sanctions. If low-return players, however, decide to contribute c_w^{min} , our setting basically allows for two reactions. First, contribute fully and sanction to equity afterwards. Second, contribute less than fully and settle for contributions that are not on the Pareto-frontier. Which of the two strategies is followed depends on whether low-return players are sufficiently weak in their aversion against inequality to accept the sanctioning to equity. More precisely, the degree of inequality that arises when high-return players contribute fully but apply sanctions to reach equity⁵⁵ afterwards has to be smaller than the maximum willingness of low-return players to accept inequality (which is determined by their c_w^{max} – see condition C7):

$$\hat{m}(2 \cdot 20 + 2 \cdot c_w^{min}) - \left(1 - \frac{m_l}{m_h}\right) < \hat{m}(2 \cdot 20 + 2 \cdot c_w^{max}) - (20 - c_w^{max}). \quad (\text{B10})$$

The equation implies a condition that specifies how high c_w^{max} has to be given a particular c_w^{min} : $c_w^{max} > \frac{20 + 2\hat{m}c_w^{min} + \left(1 - \frac{m_l}{m_h}\right)}{2\hat{m} + 1}$. In other words, the higher the c_w^{min} , the higher will the inequality be that results due to sanctioning. For this reason, the willingness to accept inequality has to be sufficiently high. Below, we provide an example illustrating the implications of this condition given the parameters of the experiment. If the condition is not met, sanctions would create an incentive for low-return players to sanction as well. One strategy to avoid this problem is to contribute less than fully and still apply sanctions to reach equity afterwards. Contributing less than fully implies lower payoffs for high-return players (as their joint return is bigger than 1) and, thus, a lower level of inequality. But even in this case, it may happen that low-return players may not be lenient enough.⁵⁶ Then, high-return players can contribute even lower amounts (e.g. zero) and not apply punishment at all. As low-return players contribute at least, c_w^{min} , contributing very low amounts can result in

⁵⁵ High return players earn $\pi_h = m_h(2 \cdot 20 + 2 \cdot c_w^{min}) - 1$. After sanctions, low-return player should earn $\pi_l = \pi_h \cdot m_l/m_h$ to achieve equity. Subtracting one from the other leads to the difference with respect to equality (as shown on the left-hand side of the inequality).

⁵⁶ Similarly as before, the respective condition is: $\hat{m}(2 \cdot c^h + 2 \cdot c_w^{min}) + (20 - c^h - 1) \left(1 - \frac{m_l}{m_h}\right) < \hat{m}(2 \cdot c^h + 2 \cdot c_w^{max}) - (20 - c_w^{max})$, where c^h indicates the high benefit players contribution that is smaller than 20.

very low levels of inequity (or even advantageous inequity). Of course, since the joint return of high-return players is higher than 1, they will only make use of this strategy if low-return players' aversion to inequality does not allow for the former strategies (of higher contributions and the application of punishment).

In summary, if the condition on p is met, contributing c_w^{min} can be rationalized as equilibrium play for low-return players. (Medium averse high-return players contribute $b_h^m(c_w^{min})$ and weakly averse ones contribute fully.) Contributing c_m^{min} can, however, either not be rationalized in equilibrium or at least leads to lower earnings for low-return players. If p is high enough, low-return players that contribute c_m^{min} would, in principle, like to reduce their contributions. The medium averse high-return type (that contributes fully) would react to insufficient contributions by sanctioning to equity. If low-return players are willing to accept that (i.e., if they are not too inequality averse), it is optimal for them to deviate to c_w^{min} . If they are not, they would “counter-sanction”, making a *unilateral* deviation potentially unprofitable, depending on how high p exactly is. Of course, low-return players contributing c_w^{min} may in any case not be the only equilibrium strategy as zero contributions of all players may also be an equilibrium outcome. ■

Thus, in a setting with only three player types (two high-return and one low-return), low return-players may have an incentive to provide insufficient contributions for more strongly inequity-averse high-return players. In other words, without communication, participants in our experiment may be unsure about the preferences of other players. If low-return players underestimate the strength of high-return players preferences, they may not contribute enough. This may in turn lead to punishment (or at least Pareto-inefficient contributions).⁵⁷ Notably, punishment will, however, be limited in our simple scenario as the latter players adjust their strategy to avoid heavy punishment and feuds. Of course, the inclusion of further players can change that as they can create an incentive for high-return players to contribute fully even though this implies feuding with the low-return players. Before discussing the more general environment, let us first look at an example.

Let us illustrate the setting with an example that follows the parametrization of our experiment. We will consider a situation in which both extreme outcomes – namely efficiency and equal payoffs – are possible. Thus, in an abuse of notation, inequity and inequality aversion to a “medium” degree refer to the extreme case of efficiency or full contribution. In other words, we assume $c_m^{min} = 20$

⁵⁷ In our theoretical environment, we implicitly assume that players are ex-post inequality/inequity averse, i.e., they care about the inequality/inequity that arises in particular realization of player types. Outcomes would change if we assumed that players were solely (or predominantly) ex-ante inequality averse, i.e., they care about the inequality in expectations. As groups in our experiment remain constant during the whole experiment, the former assumptions seems, however, more plausible for our particular setting.

and that $c_m^{max} = 20$. Similarly, weakly equity-minded individuals are willing to accept equal earnings, $c_w^{min} = 5$.

To assess what the optimal action for low-return players is, we will consider what utilities arise from different strategies. If low-return players contribute $c_l = c_w^{min} = 5$ and interact with a (compatible) weakly equity-minded high-return players, equal payoffs (without any punishment) emerge and no disutility due to inequality occurs. Thus, $u_l^m(20, 5, p = 0) = \pi_l^m(20, 5, p = 0) = 30$. In contrast, if low types contribute fully, $c = c_m^{min} = 20$, no punishment will be meted out independently of what high-return player type is met. Of course, disutility due to the inequality does occur. Since the low-return players' c_w^{max} is 20, we can infer that $\alpha \approx 3/48$ (following the proof of proposition 2): Notably, the disutility (due to inequality) associated with the efficient outcome is equal to the cost of punishment, which is a flat fee of one. Thus, $u_l(20, 20, p = 0) = 23$. If low-return players contribute 5 and meet a high-return player that is equity-minded to a “medium” degree, equal earnings of 30 arise if the latter players contribute fully. Equity-minded players are not willing to accept this and restore equity by punishing to 14.5. Notably, condition (B10) is met as weakly equality-minded players are lenient in accepting inequality.⁵⁸ Hence, they do not have incentive to punish and receive a utility of $u_l^m(20, 5, p > 0) = 13.9$. When will low-return players choose to contribute only c_w^{min} ? They do so as long as their chance of meeting high-return players who are only weakly inequity-averse is sufficiently high. Using proposition 3's condition, we can see that p has to be larger than 0.57. Then, it is optimal for weakly inequity-averse players to contribute only 5. Sometimes – with probability $1 - p$ – strongly minded high-return players are met and contributions are insufficient. Notably, feuds cannot occur in our example as (a) players are too lenient in accepting inequality and (b) medium inequity-averse high-return players have incentives to avoid such a feud even if low-return players were inclined to counter-punish. As low-return players are willing to accept quite some inequality, contributing c_m^{min} cannot be rationalized in equilibrium as deviating from it will always be profitable for a low-return player when the condition on p is met.

Complex setting and feuding. When will feuds occur? In our simple environment, high-return players (with a medium degree of *inequity* aversion) have an incentive to avoid feuds as their action solely depends on one type of inequality-averse player. With additional player types, some high-return players may, however, feud with low-return players that contribute insufficiently as long as there are (sufficiently many) other low-return players that fulfil their contribution requirement and make it worth contributing fully. Let us return to the more complex setting. Compared to the simple

⁵⁸ They would even accept a difference in payoffs of up to 24 ($= 48 - 24 = \pi_h(20, 20, p = 0) - \pi_l(20, 20, p = 0)$).

environment, a high-return player type that is *strongly* inequity-averse is added, as well as a compatible low-return player that is *weakly* inequality averse so that:⁵⁹

$$c_{w(eak)}^{max} = c_{s(trong)}^{min} \gg c_{m(edium)}^{max} = c_{m(edium)}^{min} > c_{w(eak)}^{min}.$$

Here, we assume that medium inequality-averse players are sufficiently inequality averse that they will not accept a punishment to equity after contributions have led to equal payoffs, a prerequisite for a feud (i.e. for the parameters of our experiment $c^{max} < 15$). In addition and for simplicity, we also assume that the preferences of mediumly *inequality*-averse players and strongly *inequity*-averse players are such that they are fundamentally incompatible ($c_{s(trong)}^{min} \gg c_{m(edium)}^{max}$), i.e., under any circumstances a destruction of payoff arises.⁶⁰ This simplification directly implies that the contributions of strongly *inequity*-averse players solely depend on the contributions of weakly *inequity*-averse players (that potentially might fulfil their contribution requirement).⁶¹ Thus, both new player types are likely to make high contributions as long as the strongly inequity-averse player type is sufficiently frequent that weakly inequality averse players have an incentive to fulfil the highest contribution requirement $c_{s(trong)}^{min}$. In turn, however, medium *inequity*-averse players may now – in the more complex environment – have an incentive to contribute fully even if this implies feuds with medium *inequity*-averse players that contribute insufficiently. The new low-return player type simply has to be frequent enough to make full contributions worthwhile. The following proposition follows this intuition. Table B1 illustrates the situation by showing which contribution requirement low-return types have to fulfill to avoid punishment when meeting one of the high-return player types (assuming high-return players contribute fully). Of course, low-return players know their own type but not that of the high-return players. Let q denote the probability that *low-return* players are weakly inequality-averse and $1 - q$ the probability that they are averse to a medium degree. Let $1 - r$ denote the probability that *high-return* players are strongly inequity-averse and r the probability that they are either weakly or medium averse to inequality. Abstracting

⁵⁹ Notably, one can show that simply adding another low-return player type is not sufficient. The newly added low-return player type needs to be supported by an added high-return player type.

⁶⁰ Note here, that the levels of incompatibility of preferences are slightly different when we assume that we have only two decision makers. Following a similar, but adjusted calculation as the one presented in B.3, we now get $c_{feud}^{max} \leq 10$ and $c_{feud}^{min} > 15$. In other words, as long as high-return players demand more than 15 and low-return players are only willing to contribute 10 or less, a destruction of payoffs is inevitable. If low-return players are, however, only willing to contribute 10 or less, they will also be unwilling to accept that high-return players punish to equity.

⁶¹ Of course, alternatively assuming that preferences of the two player types are incompatible ($c_s^{min} > c_m^{max}$) but do not necessarily lead to feuding could lead to qualitatively similar results if the weakly inequality-averse player type is sufficiently frequent. This would imply that strongly inequity-minded players have an incentive to contribute fully as long as the former players fulfil their requirement independently of what inequality averse players with a medium degree of aversion do.

from any strongly averse low-return players, let p denote the probability of weak aversion to inequity and $1 - p$ the probability of a medium aversion.

Table B1 - Choices of low-return players (that would not result in sanctions)

			Low return player types	
			$1-p$	p
			weak	medium
High-return player types	$l-r$	strong	c_s^{min}	X
	$l-q$	medium	c_s^{min}/c_m^{min}	c_m^{min}
	r	q	$c_s^{min}/c_m^{min}/c_w^{min}$	c_m^{min}/c_w^{min}

Notes: The table shows choices of low-return players that would not result in punishment when meeting one of the high-return player types (assuming high-return players fully contribute). The "X" indicates that aversions of the involved player types is unavoidable.

Proposition 3 (b): [Ex-Ante and Ex-Post Normative Conflict: Feuds]

If the condition of Proposition 3(a) is met, there exists a frequency of the strongly averse high-return player type, $1 - r < 1$, and a frequency of the weakly averse low-return player type, $q < 1$, such that the presence of the new player types implies the following: It can be rationalized as equilibrium play that high-return players with a medium degree of inequity aversion do not shy away from feuding with mediumly inequality-minded players.

Proof.

Under our simplifying assumption, strongly minded high-return players will always contribute fully as long as weakly minded low-return players fulfill their contribution requirement ($c_l^w = c_s^{min}$). In particular, the former players' contribution decision is not affected by low-return players with a stronger degree of inequality aversion as the preferences of both player types are assumed to be incompatible anyway.

Under which conditions will *weakly* inequality-averse players fulfill the most stringent requirement? By fulfilling the most stringent requirement and contributing c_s^{min} they can avoid any punishment and secure a utility of $u_l^w(20, c_s^{min}, p = 0)$ for the interaction with all potential high-return player types (assuming those contribute fully). Under which circumstances will contributing less than c_s^{min} be optimal? The best case scenario for contributing less is that contributing c_w^{min} will lead to a punishment-free payoff of $u_l^w(20, c_w^{min}, p = 0)$ with probability r . In other words, we look at the case where low-return players get the best possible payoff with probability r , effectively assuming that there are no medium high-return player types. Contributing c_w^{min} , however, still implies reduced utility when meeting the strongly minded high-return player, $u_l^w(b_h^s(c_s^{min}), c_s^{min}, p \geq 0)$ with probability $1 - r$, where $b_h^s(c_s^{min})$ is the strongly equity-minded players' best response to insufficient contribution. As discussed in the previous proof, they will react to insufficient contribution in the following way: they can either contribute fully and punish to equity (in case low-return players accept that) or contribute less (and potentially punish but avoid counter-punishment of low-return players). Notably, as $u_l^w(20, c_w^{min}, p = 0)$ is strictly larger than $u_l^w(20, c_s^{min}, p = 0)$ [which is in turn larger than $u_l^w(b_h^s(c_s^{min}), c_s^{min}, p \geq 0)$], there always exists a frequency of the strongly *inequity*-averse player type (with high-returns), $1 - r < 1$, such that it optimal for weakly *inequality*-averse players to fulfil the requirement of strongly *inequity*-averse players (as long as all high-return types contribute fully).

If weakly *inequality*-averse players contribute c_s^{min} , it can provide high-return players with a medium degree of *inequity* aversion with an incentive to contribute fully even if this leads to feuding with more strongly *inequity*-averse players. Contributing fully implies a payoff of $u_h^m(20, c_s^{min}, p = 0)$ with probability q when meeting the weakly minded low-return players, and a utility of zero with probability $1 - q$ assuming that low-return players with a medium degree of *inequity* aversion contribute insufficiently (and at the same time do not accept the resulting punishment to equity), i.e. feuding occurs. Alternatively, high-return players could reduce their contributions along the lines of the description of the previous proof, $\tilde{b}_h^m(c_w^{min}) < 20$, to avoid feuding with medium *inequity*-averse players, implying a utility of $u_h^m(\tilde{b}_h^m(c_w^{min}), c_w^{min}, p \geq 0)$ with probability $1 - q$. At best, if weakly *inequality*-averse players do not punish insufficient high type contributions, this could lead to a punishment-free payoff of $u_h^m(\tilde{b}_h^m(c_w^{min}), c_w^{min}, p = 0)$ with probability q . Since the joint return of high-return players is greater than 1, $u_h^m(20, c_s^{min}, p = 0) > u_h^m(\tilde{b}_h^m(c_w^{min}), c_s^{min}, p = 0)$. Thus, even if we assume the best case scenario, that weakly *inequity*-averse players do not punish insufficient contributions, a frequency of this player type exists, $q <$

1, such that fully contributing is optimal for mediumly *equity*-minded players (assuming the former typs contributes c_s^{min} .⁶² This even holds true when full contribution involve feuding.

Will medium *inequality*-averse players (with low-returns) engage in feuding? Note that we have assumed that these players are sufficiently inequality-averse that they will not accept any punishment to equity after a contribution profile of $(c_h = 20, c_l = c_w^{min})$. It is optimal for them to contribute insufficiently, i.e. c_w^{min} , as long as the frequency of weakly equity-minded players is sufficiently high relative to mediumly equity-minded players. In other words, condition (B9) has to hold. Unlike before, mediumly equity-minded players still have an incentive to contribute fully under the conditions described above and, thus, feuds will occur, implying that in condition (B9) $u_l^m(b(c_w^{min}), c_w^{min}, p \geq 0)$ will be zero. Thus, feuding can be rationalized in an equilibrium, in which all high-return player types contribute fully, weakly inequality-averse players contribute c_s^{min} and mediumly inequality-averse player contribute c_w^{min} . ■

In summary, in our complex setting, both strongly and medium *inequity*-averse players will feud with medium *inequality*-averse players as long as the probability conditions are met. Notably, the feuding of the two latter players solely arises due to the incomplete information. Put differently, if some low-return players are unsure about the degree of aversion of high-return players and underestimate it, they might contribute insufficiently (as already indicated in our simple setting). This in turn can lead to feuding, as long as high-return players still have an incentive to contribute fully.

Let us also illustrate our complex setting with a simple example that follows the parametrization of our experiment. As before, we include the two extreme solutions in our example and assume $c_s^{min} = 20$ and that $c_w^{max} = 20$. Similarly, weakly equity-minded individuals are willing to accept equal earnings, $c_w^{min} = 5$. The two medium types require (or are willing) to contribute $c_m^{min} = 10$ ($c_m^{max} = 10$).

As indicated in the previous proof, strongly equity-minded players will contribute fully as long as weakly inequality-averse players do the same. The latter player type can gain $\pi_l^w(20, 20, p = 0) = 24$ with certainty. As its c_w^{max} is 20, we can infer that $\alpha \approx 3/48$, implying $u_l^w(20, 20, p = 0) = 23$. At best, s/he can gain $u_l^w(20, 5, p = 0) = 30$ with probability r and $u_l^w(20, 5, p > 0) = 13.9$ with probability $1 - r$ (see also previous example for calculations). Thus, if $1 - r > 0.43$, this player type will have an incentive to contribute fully.

⁶² In many cases, weakly inequality-averse players may even be inclined to punish insufficient contributions of high-return players, implying that q will even be lower.

Thus, if the condition on r is met, the two newly added players in our complex setting will fully contribute. This, however, can create an incentive for the remaining players to feud with each other. Let us consider the incentives of the low-return players with a medium aversion to inequality. Here, we can abstract from their interaction with strongly equity-minded players as their preferences are not compatible anyway, implying a destruction of payoffs under any circumstances. Focusing on the two other remaining high-return player types, low-return players with a medium degree of inequality aversion can either secure themselves a punishment-free utility with certainty by contributing $c_m^{min} = 10$, $u_l^m(20, 10, p = 0) = 27$. Alternatively, they can get $u_l^m(20, 5, p = 0) = 30$ with probability p by contributing $c_w^{min} = 5$. Under the assumption that feuding will occur, the payoff will be zero with probability $1 - p$. Thus, as long as $p > 0.9$, only $c_w^{min} = 5$ will be contributed (even in case of feuding).

Finally, will the high-return players with a medium aversion against inequity shy away from feuding? Not if weakly inequality-averse players are sufficiently common. While contributing fully will, with probability $1 - q$, result in a destruction of payoffs in case medium inequality-averse players are met, with probability q , a payoff of $u_h^m(20, 20, p = 0) = 48$ occurs. Alternatively, the feud will be avoided by contributing $c_h^m = 10$ ($c_l^m = 5$). This results in high-return player earnings of $\pi_h^m(10, 5, p = 0) = 28$ and low-return player earnings of $\pi_l^m(10, 5, p = 0) = 24$, implying the same level of inequity of a contribution profile of $(c_h = 20, c_l = 10)$ and, hence, no need for punishment. Thus, with probability $1 - q$, $u_h^m(10, 5, p = 0) = 27$. Notably, in our particular example, the weakly inequality-averse player will not be happy about any contributions below 20 as we assume that $c_w^{max} = 20$, and, thus, punishes to equality, resulting in $\pi_h^m(10, 20, p > 0) = 17 [= \pi_l^w(10, 20, p > 0)]$. This implies $u_h^m(10, 20, p > 0) = 16.15$. Thus, as long as $q > 0.46$, a high-return player with a medium degree of inequity aversion will not shy away from feuding.

In summary, under incomplete information, incorrect beliefs may lead to insufficient contributions in one's group, which in turn may even lead to feuds. While this may certainly not be true under all possible group preference configurations, it seems a plausible solution in groups in which group members with strong inequality and inequity preferences interact. Of course, on top of the described feuding, incomplete information may also increase feuding that does not result from rational maximization behavior.⁶³

⁶³ Let us briefly discuss potential implications of playing a game with two (or more) rounds instead of a one-shot scenario. This creates the potential that inequality-averse low-return players would update their beliefs over time. High-return players that punish low-return players in the first round reveal that their inequity concerns are stronger than the ones expected on average. In our simple setting, low-return players that start by contributing c_w^{min} in the first round could contribute c_m^{min} from the second round onward after observing punishment in the previous round. We generally acknowledge that ex-ante normative conflict might be less persistent than ex-post normative conflict and

No conflict. So far, we have described a situation, in which there is (resolvable) ex-post normative conflict under complete information. We have shown that in this situation, uncertainty about the degree of others' aversion to inequality/inequity can lead to ex-ante normative conflict, i.e., punishment and even feuds that result under incomplete, but not complete, information, in a particular realization of the aversions against inequality and inequity in one's group. Let us finally look briefly at a situation, in which there is a chance that all players in a group are inequity-averse but they may still not reach efficiency because they fear the presence of an inequality-averse player. In other words, ex-ante normative conflict is present, but it is merely caused by the fear of ex-post normative conflict rather than its actual presence.

As before, we consider a simple, stylized situation to make the argument transparent. The setting we consider requires a particular set of assumptions. Importantly, we do not claim that ex-ante normative conflict without ex-post normative conflict is a widespread problem, but we want to make the point that it can exist. We will assume that there are four players and that *both* high-return players are equity-minded to a weak degree, i.e. they have a fairly high tolerance for a potential inequality type, c_w^{min} . Unlike earlier, we will now consider the two low-return players separately. Nature determines that the first low type, LOW1, is either *equality*-minded to a fairly strong degree with probability p , or *equity*-minded to a fairly weak degree with probability $1 - p$. For simplicity, we assume for the former case that the low-return player has the same level of inequity tolerance as the two high-return players, c_w^{min} . Similarly, for the latter case, we assume that the LOW1's inequality tolerance is given by c_s^{max} and we assume that $c_s^{max} = c_w^{min}$, i.e., aversions against equality and equity are exactly compatible with each other. The second low type, LOW2, is either *equality*-minded to a fairly strong degree with probability q or (very) strongly *equity*-minded with probability $1 - q$. For the former case, we assume that the degree of inequality aversion is the same as for LOW1. For the latter case, we assume that the aversion against inequity is fairly strong and –for simplicity – we actually assume that LOW2 only accepts efficient payoffs $c_{eff}^{min} = 20$. Overall, it holds that the varying degrees of inequity and inequality-aversion match in the following way:

$$20 = c_{eff}^{min} \gg c_s^{max} = c_w^{min}.$$

disappear over time. Nonetheless, one has to bear in mind that even theoretically, the update process can take more than one round when one – more realistically – considers a whole range of degrees of inequity aversion. If low-return players who contributed a low amount are punished in the first round, they might only increase their contribution slightly if they still think that there is a large fraction of high-return players with an intermediate degree of inequity aversion. Similarly, our analysis abstracts from differences in c^{min} (c^{max}) among low-return (high-return) players that might updating more difficult.

Let us first consider what these (in-)equalities imply for case of complete information. We will focus again on symmetric equilibria on the Pareto-frontier: If both low types, LOW1 and LOW2, (as well as the high-return players) are equity-minded, LOW2's strictness in his or her aversion against inequity ensures that an equilibrium exists in which everyone contributes fully.⁶⁴ Similarly, if both low types are equality-minded, the high return players' aversion against inequity ensures an equilibrium in which low types will contribute c_m^{min} (while high types contribute fully). Notably, if LOW2 is equality-minded and LOW1 is equity-minded, the group can still "coordinate" on the same outcome as long as LOW1 is not too averse against *advantageous* inequity as LOW1 will be ahead in terms of equity in this equilibrium.⁶⁵ Only, if LOW1 is equality-minded and LOW2 is equity-minded preferences are not reconcilable since LOW2 is very strong in his or her aversion against inequity.

Notably, with incomplete or imperfect information, a situation may arise in which all players are equity-minded and efficiency is obtainable but LOW1 does not contribute fully due to his or her fear that LOW2 is actually equality-minded. What are players' incentives under imperfect information assuming the (in-)equalities as outlined above? An *equality*-minded LOW1 knows that s/he will not be able to agree on a contribution schedule with the (strictly) equity-minded LOW2 and thus will contribute c_w^{min} as long as equality-minded LOW2 does the same (and high-return players contribute fully). As long as the likelihood of LOW1 being equality-minded is *sufficiently* high, LOW2 has an incentive to contribute c_w^{min} irrespective of *equity*-minded LOW1's decision.⁶⁶ Crucially, however, if the latter player type decides to contribute fully, this would imply an unacceptable increase in inequality (as high-return players benefit more from contribution than low-return ones) that would require *equality*-minded LOW2 to sanction *equity*-minded LOW1. Of course, *equity*-minded LOW2 will always want LOW1 to contribute fully. This implies that *equity*-minded LOW1 has to choose between contributing fully for the case that LOW2 is equity-minded or c_w^{min} for the case that LOW2 is equality-minded. If the probability of the latter is sufficiently high, s/he will choose c_w^{min} . Anticipating this, equity-minded LOW2's best response depends on his or her willingness to accept *advantageous* inequity. While contributing zero can still avoid

⁶⁴ If all players' aversion against *advantageous* inequity is high enough, all group members will even contribute fully voluntarily. But even if this is not true, LOW2's (strict) aversion against disadvantageous inequity ensures that efficient contributions are reached.

⁶⁵ Additionally, equilibria are possible in which both high types contribute fully, LOW1 contributes more than c_m^{min} , potentially even fully, and LOW2 contributes less than c_w^{min} . We will see below that these equilibria are not attractive to LOW2 in a setting with incomplete information, as long as LOW1's probability of being equality-minded is high enough.

⁶⁶ Equality-minded LOW2 can either interact with equity-minded or with equality-minded LOW1. While asymmetric equilibria exist in the complete information case for the former interaction, the greater the likelihood that LOW1 is equality-minded the stronger is the incentive to choose c_w^{min} .

disadvantageous inequity this is not true on the advantageous side. If the strict aversion to inequity includes advantageous inequity, LOW2 can contribute fully and sanction LOW1 and thus fully restore equity. Of course, LOW1 will not “counter-punish” as s/he shares the fairness benchmark. Overall, this implies, that a situation (or realization) arises in which all players are actually equity-minded but LOW1 contributes insufficiently due to his fear of LOW2 being equality-minded. Of course, if all subjects are actually equity-minded, it should be relatively easy to update one’s belief and adjust one’s behaviour over time. Thus, without actual ex-post normative conflict, not too much punishment would be observed.

B.5 Miscellaneous results

In the section, we discuss two remaining topics. First, we briefly consider the homogenous case. Second, we discuss what happens if high-return players are inequality and low-return players are inequity-averse.

Homogenous returns. For completeness and comparison, we will briefly discuss the implications of inequality aversion in the homogenous setting. Naturally, inequality aversion coincides with inequity aversion in the homogenous setting. In other words, there is no *ex-post* normative conflict in such an environment. In addition, since there is no heterogeneity, it seems a reasonable assumption that even without communication, one can assume complete information, so that there is no *ex-ante* normative conflict either.⁶⁷ Thus, proposition 0 simply considers the public good environment featuring homogenous returns, with and without punishment. It indicates that, compared to the heterogeneous setting, fairly similar results emerge, at least if we compare it to the case of no normative conflict.

Proposition 0. (a) *No punishment:* If $\beta_i \leq \frac{n-1}{n-3}(1-m)$ for all i ,⁶⁸ then

1. *there are no equilibria with asymmetric contributions or payoffs,*
2. *the vector of zero contributions is a Nash equilibrium, and*
 - a. *if $\beta_i < 1-m$ for some $i \in \{1, \dots, n\}$, then there are no other equilibria,*

⁶⁷ Similarly as in C.4, agents could suffer from uncertainty about the degree of inequality aversion that other players have. They will, however, not fear that other players might be inequity-averse. While the former uncertainty could still imply that some players contribute too little (which could in turn result in punishment), it is clear that no counter-punishment can occur insofar as all agents share the same motivation to contribute or to free-ride.

⁶⁸ It also holds that there are no equilibria with asymmetric contributions if $n \leq 3$.

b. if $\beta_i \geq 1 - m$ for all $i \in \{1, \dots, n\}$, then any vector of identical contributions is a Nash equilibrium. This includes both the equality rule as well as the efficiency rule as they coincide.

(b) **With punishment:** Suppose that there is one high-return player that is sufficiently inequality-averse, i.e., his or her preferences obey $\beta_i \geq 1 - m$ and $\alpha_i > n + 1$. In addition, suppose that all the other players are selfish, then the following strategies that describe the players behaviour on and off the equilibrium path form a subgame perfect equilibrium:

- In the contribution stage, each player contributes $c_i = \hat{c} \in [0, 20]$.
- If each player does so, no player is sanctioned in the second stage. If a player j deviates by contributing $\hat{c} - \Delta$, $\Delta > 0$, then the sufficiently inequality-averse player i – the enforcer – chooses $p_{ij} = \Delta + 1$ and $p_{ik} = 1$ for $k \neq j$.

Proof. Part (a). Consider any vector $\mathbf{c}$ with non-identical contributions. Let agent i be one of the individuals with the largest contributions in $\mathbf{c}$ such that there are no contributions larger than c_i . Since contributions are non-identical, there exist at least one individual j with a smaller contribution than c_i . Suppose that agent i deviates and offers slightly less than c_i : say he or she offers $c_i - \Delta$, where Δ is such that $c_i - \Delta$ is strictly greater than the second highest contribution in $\mathbf{c}$. By doing so, i increases his or her own private payoff by $(1 - m)\Delta$. In addition, i increases his or her utility by reducing the difference to players with higher earnings. Since $\mathbf{c}$ leads to non-identical payoffs, there is at least one such player, implying that the minimal gain is $\frac{\alpha_i}{n-1}\Delta$ (which is larger than $\frac{\beta_i}{n-1}\Delta$). In addition, however, the deviation is associated with a utility loss as it increases inequality with people of the same payoff. There are at most $n - 2$ such players, resulting in a maximum loss of $\frac{\beta}{n-1}[n - 2]$. Thus, deviating is profitable as long as

$$\left((1 - m) + \frac{\alpha_i}{n - 1} \right) \Delta > \frac{\beta_i}{n - 1} [n - 2] \Delta.$$

Since $\beta_i \leq \alpha_i$, this condition holds if $\beta_i \leq \frac{n-1}{n-3}(1 - m)$. Hence, a sufficient condition for agent i to be better off in case of contributing $c_i - \delta$ is $\beta_i \leq \frac{n-1}{n-3}(1 - m)$, which is true by assumption. Thus, there are no equilibria with asymmetric contributions (or payoffs). It can be shown in a straightforward manner that a similar result holds for $n \leq 3$.

We will now derive the conditions under which symmetric-contribution equilibria exist. Note, however, first that the vector of zero contributions, $(0, \dots, 0)$, is a Nash equilibrium. Indeed, in this case, the utility of any agent is simply 20 and any unilateral deviation necessarily involving a

strictly positive contribution, e.g., $c_i = \varepsilon > 0$, results in $u_i(c_i, 0) = 20 - \varepsilon + m\varepsilon - \alpha_i\varepsilon$, which is smaller than 20 because $m < 1$.

Now let $\hat{c}$ be a vector of identical contributions, $\hat{c}_i \geq 0$, for all $i = 1, \dots, n$. The utility of some agent i is $u_i(\hat{c}) = 20 + (nm - 1)\hat{c}$, whereas in the case of investing less than $\hat{c}$ – say, $\hat{c} - \delta, \delta \in (0, \hat{c}]$ – agents i 's utility becomes $u_i(\hat{c} - \delta, \hat{c}) = 20 - \hat{c} + \delta + m(n\hat{c} - \delta) - \beta_i\delta = u_i(\hat{c}, \hat{c}) + \delta(1 - m - \beta_i)$. This implies that if there is an agent with $\beta_i < 1 - m$, then he or she is better off by contributing less than $\hat{c}$, and therefore $\hat{c}$ is not a Nash equilibrium. On the other hand, if $\beta_i \geq 1 - m$, for all i , then the last equality implies that nobody is better off by deviating down. Similarly as for the vector of zero contributions, $(0, \dots, 0)$, it can be shown that deviating up cannot increase utility.

Part (b). Note first, that if everyone contributes according to the schedule $\hat{c}$ no punishment will take place. The enforcer does not suffer any disutility due to inequality and, thus, has no reason to punish. Suppose that a player j deviates by contributing $\hat{c} - \delta$. If all other players contribute $\hat{c}$ and the enforcer sticks to the punishment strategy, then the deviator j gets the same monetary payoff as the enforcer i . In this case, monetary payoffs are given by

$$\pi_i = 20 - \hat{c} + m[n\hat{c} - \delta] - 1$$

$$\pi_j = 20 - (\hat{c} - \delta) + m[n\hat{c} - \delta] - \delta - 1 = 20 - \bar{c} + m[n\hat{c} - \delta] - 1 = \pi_i.$$

Notably, this implies that the “gain” of the deviation of $(1 - m)\delta$ is overcompensated by the punishment of $\delta + 1$, implying a payoff loss of $\delta + 1 - (1 - m)\delta = 1 + m\delta$. Thus, j does not gain from deviating. We now have to check whether punishment is credible, whether the enforcer is really better off from punishing compared to not punishing. Thus $u_i(c, p > 0) > u_i(c, p = 0)$. While punishing involves a (normalized) cost of 1, not punishing increases advantageous disutility with respect to the deviator by δ :

$$u_i(\hat{c}, p > 0) = 20 - \hat{c} + m[n\hat{c} - \delta] - 1 > 20 - \hat{c} + m[n\hat{c} - \delta] - \frac{\alpha}{n-1}(\delta) = u_i(\hat{c}, p = 0)$$

This inequality holds if $1 < \frac{\alpha}{n-1}\delta$. Since the smallest possible deviation is one unit in the experiment, the inequality holds for any possible deviation if $\alpha > n - 1$. In addition, the conditional enforcer cannot gain from deviating in the contribution stages as long as $1 - m < \beta$, for the same reason shown in part (a) of the proof. Of course, it is easy to show that deviating by choosing $c_i > \hat{c}$ cannot be profitable for any player. ■

Does introducing heterogeneity make cooperation more difficult? Looking at all four propositions jointly seems to provide a clear answer: Only if one assumes that the fact that high-

return players *earn* their advantage convinces low-return players to accept the equity norm, should one expect cooperation to be similarly effective as in the homogeneous case. If, on the other hand, one assumes that the benchmark of equality is so dominant that even high-return players follow it, cooperation rates should still be similarly high as in the homogeneous case but payoffs would obviously be well below the social optimum.

Of course, there are many reasons (e.g., self-serving biases) to believe that neither of the two cases is particularly realistic. If one assumes that conflicting normative views exist in many of our groups, i.e. that ex-post normative conflict is present, cooperation seems much more complicated in the heterogeneous case than in the homogeneous case. First, without punishment no (strictly positive) contribution schedule can be rationalized in equilibrium, unlike in the homogeneous case (Prop. 1). Second, with punishment, the possibility of total destruction of payoffs exists if the aversions to inequality and inequity are incompatible and strong (Prop. 2). But even if players' preferences allow for a compromise, it may not be certain that this compromise would be reached. Incomplete information (or incorrect beliefs) about the degree of others' inequity or inequality aversion could easily lead to equilibria that involve considerable punishment (Prop. 3). While players could learn the correct degree over time, it is conceivable that initial off-equilibrium punishment further harms cooperation (even though our model does not formally predict such an effect).

Counter-intuitive normative conflict. In the main text, we assumed that if conflicting normative views coexist, high-return players are inequity-averse while low-return players are inequality-averse. This would be in line with self-serving biases. In addition, our experimental results seem to be consistent with the implications of our framework. In particular, compromises between efficiency and equality that we observe in the covenants groups make are well predicted by our model. Here, we very briefly discuss the case where high-return players who are inequality-averse coexist with low-return players that are inequity-averse.

Low-return *inequity*-averse players have an inclination to contribute fully (at least if they expect others to do the same or are able to punish those that do not contribute fully). Thus, consider a situation in which all players contribute 20. (Sufficiently) *inequality*-averse high-return players incur a potentially substantial utility cost in this situation due to advantageous inequality. They can, however, not do much about it. In principle, they would like to contribute even more to reduce inequality. However, this is not possible. In addition, unilaterally contributing less would only amplify inequality. And even though punishment is cheap, it is never optimal for a high type to destroy one's own payoff to reduce inequality, as high types are ahead and $\beta_i \leq 1$.

Bearing these considerations in mind, one can show that – without punishment – groups that only consist of sufficiently inequality-averse high-return players and sufficiently inequity-averse low-return players can apply the efficiency rule in equilibrium. Of course, if punishment is available, free-riders can be disciplined to contribute fully as well.

Appendix C – Communication data

We asked three coders who were unaware of our research hypotheses to independently code the content of the messages exchanged between subjects in the communication stages of our three treatments. In the main text, we have made use of one of the classification criteria extensively: *covenants*. Before outlining the coding criteria in more detail, let us emphasize at the beginning that we analyzed the coding reliability of the communication data and found that we generally have a good inter-rater reliability and that our analysis suggest that we can safely use the classified communication data in our analysis. For the classification of an (explicit) covenant – which is the main coding variable used in the main text of our paper – the average pairwise consistency of the three coders is 80%. In addition, even when taking into account that agreement among coders can arise just by chance, we reach a level of agreement that has been classified as “substantial” (Landis and Koch 1977): The average pairwise Cohen’s Kappa is 0.61. We can clearly reject the hypothesis that agreement between coders has arisen by mere chance. Finally, we do not just look at inter-rater consensus but also consistency and calculate Cronbach’s alpha. For the covenant-classification, the alpha is 0.83, implying good consistency.⁶⁹

For each group *and* each communication stage, the coders first indicated which rules were discussed. For the purpose of the classification, a rule had to specify what both high and low types should contribute. In particular, each coder classified whether groups discussed the *efficiency* rule ($c_{high} = 20, c_{low} = 20$), the *equality* rule ($c_{high} = 20, c_{low} = 5$), and a *compromise* between the two.⁷⁰

⁶⁹ As outlined below, coders also rated the level of disagreement in a group. Since this concept is vaguer than a covenant, the measures of inter-rater reliability are, unsurprisingly, somewhat lower but still acceptable: The average pairwise consistency is 73%. The average Cohen’s Kappa is 47% (indicating that the level of agreement is moderate while we can clearly reject the hypothesis that agreement has arisen by mere chance). Finally, Cronbach’s alpha is 0.77. We use the level of disagreement to define an implicit covenant in a conservative way, imposing an additional level of agreement, as outlined below

⁷⁰ More precisely, we also allowed for a rule category in which specific contributions for each type were suggested, but the contribution of the high type was not the maximum of 20. Since these types of rules were not very common

In addition, coders classified whether *all four* group members *actively* agreed to one particular rule. This is what we call an *explicit covenant*. We first followed a strict classification – active agreement of all four group members – to avoid the ambiguity of a less stringent classification. For cases in which groups do not reach such an agreement, the coders rated the level of disagreement (or hostility) that was present in the discussion, on a scale from 0 to 4. A low level of disagreement is indicative of a mutual understanding, as we asked coders to rate a group “in which the fourth person just fails to agree to a contribution rule because time has run out” as 0. In contrast, we suggested rating “a group that discusses alternative rules and cannot agree on one rule but in which players remain polite during the whole discussion” as a 2. For a classification of 3 or 4, we required that group members threaten to punish others (“We all leave with nothing if low types do not contribute 20”), insulted one another (“You are an *****”) or behaved in similarly hostile manner. Following a conservative approach, we say that a group has established an *implicit* covenant if at least two coders assigned the lowest possible level of disagreement to the group, 0, and no coder assigned a level greater than 1. We say that groups reach a *covenant* if they either (a) reach an *explicit* covenant, or (b) an implicit covenant.

In addition, coders classified whether the *same-type chat box* was used and, if so, whether the two players discussed (a) punishing the other type players for deviations and (b) breaking promises made in the group chat box. Since most of our criteria are binary, we aggregated the data of the tree coders by the majority rule, i.e. a communication session fulfills a criterion when at least two coders indicated that it did. The level of hostility was only coded if a group did not establish an explicit covenant.

Table C1 shows the likelihood with which one of the three rules is discussed in a particular communication session of each group, separately for each treatment. In addition, it shows whether these percentages are different between treatments. The table provides an indication that compromise rules have a higher likelihood of being discussed when communication is introduced early in *ComEarly* in *ComOnly* compared to *ComLate*. This is broadly in line with result 7.

Although the same-type chatbox is used fairly frequently in all three treatments (*ComLate*: 54%, *ComEarly* 66%, *ComOnly* 66%), it is not often used to discuss punishment (*ComLate*: 13%, *ComEarly* 13%, *ComOnly* 16%) or deviations from promises made in the all-group chatbox (*ComLate*: 16%, *ComEarly* 5%, *ComOnly* 35%). Moreover, we generally do not observe

and did not represent inefficient versions of the efficiency or equality rules, we included them in the compromise rule category. Unspecific comments such as “Let’s all contribute at least something” or suggestions only determining one type’s contribution “The high types should contribute x” were not classified as a discussion of a specific rule.

significant differences between treatments. In *ComOnly*, deviations are discussed more frequently than in *ComEarly* ($p = 0.001$) or *ComLate* ($p = 0.022$) in line with the finding that subjects deviate more often from covenants when the threat of punishment is not available. Moreover, correlation coefficients between using the chatbox or between using it to discuss punishment or deviations on one hand, and contributions or punishment on the other, are usually fairly low and often insignificant ($|\rho| = 0.03 - 0.26$). Overall, this suggests that the same-type chatbox may not be that important and that treatment differences are not driven by patterns in the use of the same-type chatbox.

Table C1– Discussion of Rules and Treatments

Percentages			
	Efficiency	Equality	Compromise
<i>Treatments</i>			
<i>ComLate</i>	22.2	33.3	29.8
<i>ComEarly</i>	27.4	21.5	49.0
<i>ComOnly</i>	23.5	21.5	46.0
<i>Statistical Comparison: p-values</i>			
<i>ComLate</i> vs. <i>ComEarly</i>	0.620	0.271	0.085
<i>ComLate</i> vs. <i>ComOnly</i>	0.901	0.371	0.161
<i>ComEarly</i> vs. <i>ComOnly</i>	0.772	0.673	0.751

Notes: The p-values are derived from a Mann-Whitney test and are two-tailed.

Appendix D – Analysis of coordination problems

As discussed in Appendix B, even if groups resolve normative conflict, they face a coordination problem. Groups that intend to equalize earnings could coordinate on different contribution schedules such as $(c_h = 20, c_l = 5)$, $(c_h = 16, c_l = 4)$, $(c_h = 12, c_l = 3)$, etc. Of course, the schedule in which high-return players contribute fully is Pareto-dominant. In this section, we provide evidence that groups seem not to have problems coordinating on the dominant outcome, at least if they can communicate. Without communication, a coordination problem may exist although it is difficult to disentangle it from the effects of ex-ante normative conflict (i.e. the uncertainty individuals face about the extent to which other agents may care about the different normative criteria when not being able to communicate).

With communication (in particular in Part 1 of *ComEarly*), groups that establish a covenant follow this covenant over the three periods after a communication stage with a fairly high probability of 81%.⁷¹ In addition, groups almost entirely discuss (and agree on) contribution rules that involve high-return players contributing 20. Only one group discusses (and follows) a compromise rule that involves high-return players contributing less than 20. When we do not include the first communication round, in which fewer covenants are established and the normative conflict is not entirely resolved in all groups, in our analysis, in 14 out of 16 groups high-return players consistently contribute 20 in periods 4-9 of Part 1 in *ComEarly*. Overall, this evidence leads us to conclude that there is almost no coordination problem if it is possible to communicate.

Without communication, in particular in Part 1 of *ComLate*, a very different picture emerges; full contributions of high-return players are much less likely. In periods 4-9 of Part 1 of *ComLate*, only in 5 out of 18 groups do high-return players consistently contribute 20. In addition, the overall

⁷¹ In the first communication stage especially, some groups agree on a contribution rule, but do not follow through. The most common pattern is that groups agree on a compromise rule (involving high contributions from high-return players) and low-return players give somewhat less than promised.

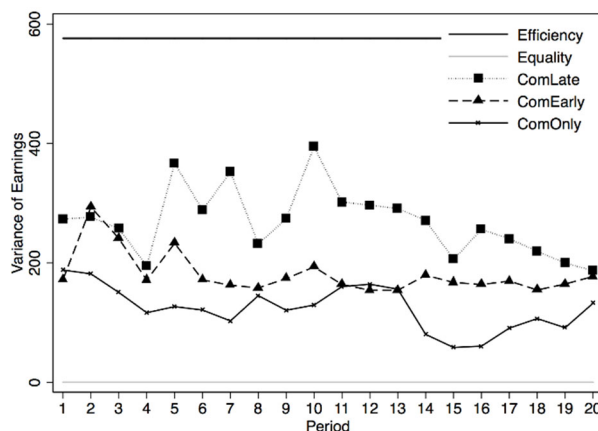
likelihood that groups follow a rule consistently over three consecutive periods is only 13%, instead of 81% when communication is possible and groups agree on a covenant. Of course, this strong difference between *ComLate* and *ComEarly* is in all likelihood not only driven by the fact that it may be more difficult to coordinate on Pareto-superior equilibria without communication. As we argue in our new theoretical framework, players may generally disagree about which normative criteria to follow when returns from the public good are heterogeneous (what we call *ex-post* normative conflict). On top of that, players face uncertainty whether and to what extent other players actually favor certain normative views (what we call *ex-ante* normative conflict), especially without being able to communicate. There are, however, at least some indications that also the “pure” coordination problem may matter. First, those few groups that do manage to follow contribution rules consistently without communication in Part 1 of *ComLate* mostly follow the efficiency rule, for which the problem of multiplicity may arguably be less severe. In addition, only in 2 out of 18 groups do high-return players start with a contribution of 20 in the first period of *ComLate* (while this is true for 12 out of 16 groups in *ComEarly*). Of course, high-return players’ initial behavior could also be explained as more cautious behavior in the presence of (ex-ante) normative conflict.

While we cannot clearly disentangle the effects of (ex-ante) normative conflict and a problem in coordinating on Pareto-superior equilibria in the absence of communication, we believe for a number of reasons that the former problem is the more fundamental one (though it might well be magnified to some extent by the latter one). First, the instructions provide subjects both with the equality and with efficiency rules, highlighting the Pareto-superior solution. Second, empirically, we do not observe groups in *ComLate* to coordinate on Pareto-inferior equilibria. Third, when communication is introduced in *ComLate* in part 2, we find no evidence that subjects explain prior cooperation failure in their messages by a mere problem of coordinating on a Pareto-superior contribution schedule. Finally, after communication is finally introduced in *ComLate*, a number of groups still sanction heavily. This may be difficult to explain if we assume that groups may actually have agreed on a particular fairness principle (equity or equality) from the beginning but simply failed to coordinate on the respective Pareto-dominant contribution schedule.

Appendix E – Analysis of variance of earnings with groups

As already indicated in the main text, concerns for reducing inequalities in earnings appear to play a significant role and lead groups away from efficiency. Figure D1 depicts the variance of earnings between players within each period. This variance is on average much smaller than under full efficiency in all treatments. In addition, there is some evidence that groups experience relatively more inequality in *ComLate* compared to *ComEarly* and *ComOnly* (Mann-Whitney, two-tailed, comparison with *ComEarly*, $p = 0.103$; *ComOnly*, $p = 0.041$). The fact that observed inequality is lower than under the efficiency rule, coupled with the tendency to adhere to covenants that specify equality or compromise between equality and efficiency, suggests that groups to some extent sacrifice average earnings to lower within-group inequality. The observation that compromise rules are less likely in *ComLate* than in *ComEarly* provides additional evidence that EANC can have an effect on how EPNC is resolved after communication is introduced in a later stage, as already observed in the previous section with respect to punishment expenditures. Finally, Table 2 also indicates that without punishment, groups may frequently agree on a covenant but have problems to follow through on them as they have no (strong) disciplinary instrument at their hands.

Figure D1 – Variance of Earnings within Groups



Appendix F – Instructions

This appendix contains the instructions for the *ComLate* treatment that were used in Tilburg. The instructions used in Abu Dhabi differed only with respect to how ECUs were converted to the local currency. In addition, the instructions for *ComEarly* and *ComOnly* were very similar. For *ComEarly*, the description of the communication possibilities was already introduced in part 2.⁷² For *ComOnly*, punishment stages were not discussed.

Instructions – Part 1

General information

You are now taking part in an economic experiment. If you read the following instructions carefully, you can, depending on your decisions, earn a considerable amount of money. It is therefore important that you take your time to understand the instructions.

The instructions that we have distributed to you are for your private information. Please do not communicate with the other participants during the experiment. Should you have any questions please raise your hand.

⁷² For the reader's convenience, we talk of parts 0, 1, and 2 in the main text while the instructions used part 1, 2, and 3.

During the experiment we shall not speak of Euros, but of Experimental Currency Units (ECU). Your entire earnings will be calculated in ECUs. At the end of the experiment the total amount of ECUs you have earned will be converted to Euros at the rate of **1 ECU = 0.045 EUR** and will be immediately paid to you in cash. In addition, we will give you a show-up fee of **EUR 5** (i.e. ~ 111 ECU).

At the beginning of the experiment, the participants will be randomly divided into groups of four. You will therefore be in a group with 3 other participants. **The composition of each group will remain the same throughout the experiment.**

The experiment is divided into three parts. Here, we explain the first part of the experiment. Once the first part is finished, you will receive detailed information about the second part of the experiment. After the second part is finished, you will receive detailed information about the third part.

The task

In the first part, all participants will perform a task. The task is the same for everyone. **You will be presented with a number of words and your task is to encode these words by substituting the letters of the alphabet with numbers using Table 1 on page 4.** The task decision screen is seen in Figure 1.

Example: You are given the word FLAT. The letters in Table 1 show that F=6, L=3, A=8, and T=19.

Once you encode a word correctly, the computer will prompt you with another word to encode. Once you encode that word, you will be given another word and so on. **This process will continue for 10 minutes** (600 seconds). All group members will be given the same words to encode in the same sequence.

Period
1 of 4
Remaining time (sec): 8

STAGE 1

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
8	12	14	10	9	6	24	22	7	5	11	3	18	1	21	16	23	2	13	19	25	4	26	17	20	15

The word you are now encoding is number 1

WORD: S P O R T
CODE:

Tips:
- When a new word appears, if there are already numbers in the boxes, they may be incorrect. You should check them and replace them with the correct ones if necessary.
- Use TAB on the keyboard to switch to the next box quickly.
- After filling codes, press "OK" button to verify the code and proceed to the next word.

OK

Figure 1

The purpose of the task

The relative performance of each individual in the task will influence his or her earnings in the next parts of the experiment. This can happen in the following way. In the second part, each participant will be given an endowment (20 ECU) and will be asked to divide it between a private and a public account. There will be two types of players in your group. We will refer to them as “Type A” and “Type B”. Of the four individuals in your group, two will be of type A and two of type B.

The difference between the two types of players is the return they get from the public account. In particular, **Type-A players will have a higher return from the public account than Type-B players**. Both types will have the same return from the private account. Information about the exact returns will be given in the second part.

The second part will consist of 10 periods.

Relative Performance in Part 1 and Type Allocation

The allocation of types depends on the relative performance of the individuals in your group. At the end of the Part 1, the computer will rank the members of the group based on the number of words they encoded. **The two group members that rank first and second will be assigned the role of Type-A players. That is, the two group members with the highest number of encoded words will be assigned the role of Type-A players.** The two group members that rank third and

fourth will be assigned the role of Type-B players. If two or more participants tie, the computer will determine their type randomly.

Once Part 1 is over, you will be informed as to whether you are of Type A or Type B and receive new instructions about the second part. You will not be informed about the precise number of words encoded by each group member until the end of the experiment. This will be done using a screen as the one in Figure 2. Note that certain information has been purposefully omitted from the figure. Note also that the number of words encoded is used for the purposes of this example and should not be taken as evidence of the number of words that one can or should encode. We expect participants to be able to encode more words than the ones in Figure 2 in the allocated time.

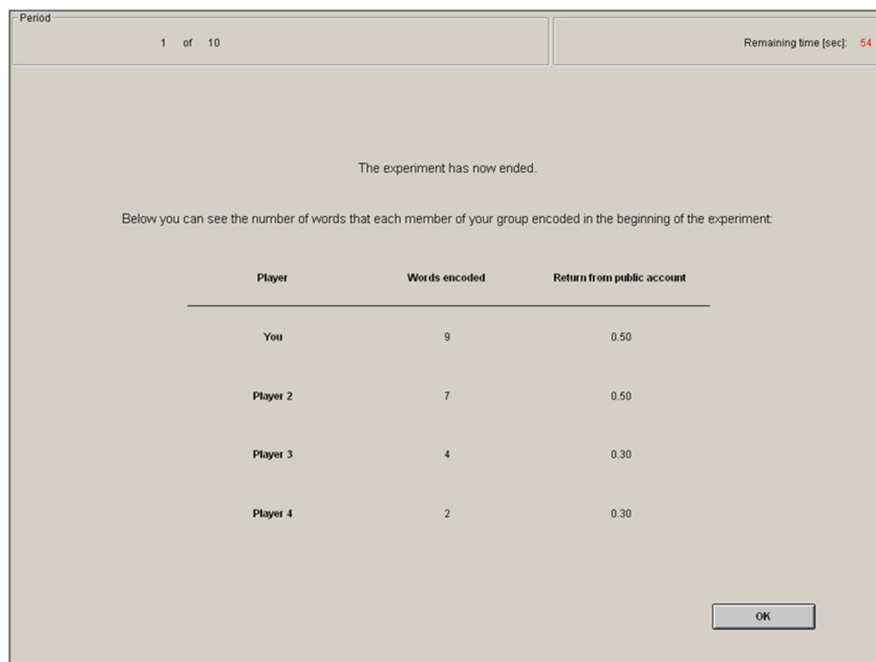


Figure 2

Table 1

Letters	Numbers
A	8
B	12
C	14
D	10
E	9
F	6
G	24

H	22
I	7
J	5
K	11
L	3
M	18
N	1
O	21
P	16
Q	23
R	2
S	13
T	19
U	25
V	4
W	26
X	17
Y	20
Z	15

Control Questions

Please answer the following questions. If you have any questions or have answered all questions, please raise your hand and one of the experimenters will come to you. All participants must fill answer the questions below before the experiment can begin.

1. What does the task in the first part determine?

.....

2. Consider the following example. Players 1, 2, 3, and 4 encode 5, 10, 12, and 4 words, respectively. In the boxes below tick the player(s) who will be Type-A player in part 2.

- ☐ Player 1
- ☐ Player 2
- ☐ Player 3
- ☐ Player 4

3. Please state whether the following sentences are true or false.

a. Type-A players will have a higher return from the private account in Part 2.

- ☐ True
- ☐ False

b. Type-B players will have a higher return from the public account in Part 2.

- ☐ True
- ☐ False

Instructions – Part 2

Recall that at the beginning of the experiment, you were randomly divided into groups of four. **The composition of your groups for this part of the experiment is the same as in Part 1 and will remain the same throughout the remainder of the experiment.**

This part of the experiment will last 10 periods. In the beginning of the experiment, each participant in your group will be randomly given a number for identification (i.e. Player 1, Player 2, Player 3, and Player 4). **Each participant will keep his/her identification number throughout the experiment.** This means that if, for example, you are assigned the role of Player 3 at the beginning of the experiment you will continue to act as Player 3 in future periods. Further, since the group composition remains the same throughout the experiment, the participant assigned the role of Player 1 in the first period will be the same as the participant assigned to Player 1 in all future periods. The same applies for Players 2 and 4.

Based on the number of words that you and the members of your group encoded, you will be assigned as either “Type A” or “Type B”. **Your type will remain the same throughout the experiment and will influence the value that you receive from the public account as explained below.**

Once the experiment is over the identities of each participant will be kept anonymous. You will be paid in private and at no point will your group or player number be revealed.

No one will know who was in their group or what actions were carried out by each individual.

Each of the 10 periods is divided into a number of stages. Next, we provide details for each stage:

The first stage (The Contribution Stage)

At the beginning of each of the 10 periods, each participant will receive 20 ECU. In the following, we shall refer to this amount as the “endowment”. Your task in the first stage is to decide how to use your endowment. **You have to decide how many of the 20 ECUs you want to contribute to a public account (from 0 to 20) and how many of them to keep for yourself.** You will be able to make your decision by using a screen as the one in Figure 1 (shown for a Player of Type A). The consequences of your decision are explained in detail on the next page.

Once all the players have chosen their contribution to the public account you will be informed about the group’s total contribution, your income from the public account and your payoff in this period through a screen as the one seen in Figure 2. Note that all numbers seen in the figures throughout this set of instructions are used only for illustrative purposes and should NOT be taken as a guide for action.

The screenshot shows a software interface for a game. At the top, there are two boxes: the left one is labeled 'Period' and contains the number '1'; the right one is labeled 'Remaining time [sec:]' and contains the number '77'. The main area of the screen has a light gray background and contains the following text: 'You have 20 ECU.', 'You can contribute any number of these ECU to the public account.', 'Your return from the public account will be: 0.60', and 'Please enter your contribution to the public account:'. Below the last line of text is a blue rectangular input field. In the bottom right corner of the interface is a red rectangular button with the text 'OK' in white.

Figure 1 – Contribution stage (for Type-A players)

Your earnings in each period are calculated using a formula that differs based on whether you are a Type-A player or a Type-B player. The earnings of a **Type-A player** are calculated using the following formula. (If you have any difficulties do not hesitate to ask us.)

$$\text{Earnings at stage 1} = \frac{\text{Endowment of ECUs} - \text{Your contribution to the Public Account} + 0.6 * \text{Total contribution to the Public Account}}{}$$

The earnings of a **Type-B player** are calculated using the following formula:

$$\text{Earnings at stage 1} = \frac{\text{Endowment of ECUs} - \text{Your contribution to the Public Account} + 0.3 * \text{Total contribution to the Public Account}}{}$$

This formula shows that your earnings at the end of the first stage consist of two parts:

- 1) The ECUs which you have kept for yourself (endowment – contribution)
- 2) The income from the public account, which equals 60% of the group’s total contribution if you are of Type A and 30% of group’s total contributions if you are of Type B.

Period 1	Remaining time [sec]: 28														
<table style="margin: auto;"> <tr> <td>Your contribution:</td> <td>5</td> </tr> <tr> <td>Total contribution:</td> <td>28</td> </tr> <tr> <td colspan="2"> </td> </tr> <tr> <td>Money kept:</td> <td>15</td> </tr> <tr> <td>Your income from the public account:</td> <td>16.80</td> </tr> <tr> <td colspan="2"> </td> </tr> <tr> <td>Your earnings from the contribution stage:</td> <td>31.80</td> </tr> </table>		Your contribution:	5	Total contribution:	28			Money kept:	15	Your income from the public account:	16.80			Your earnings from the contribution stage:	31.80
Your contribution:	5														
Total contribution:	28														
Money kept:	15														
Your income from the public account:	16.80														
Your earnings from the contribution stage:	31.80														
OK															

Figure 2 – Feedback screen after Contribution stage

Example: Suppose the sum of the contributions of all group members are 60 ECUs. In this case, a Type-A member of the group receives an income from the public account of:

$0.6 \times 60 = 36$ ECUs. A Type-B member of the group receives an income from the public account of: $0.3 \times 60 = 18$ ECUs. If the total contribution to the public account is 9 points, then Type-A members receive $0.6 \times 9 = 5.4$ ECUs from the public account while Type-B members receive $0.3 \times 9 = 2.7$ ECUs from the public account.

You always have the option of keeping the ECUs for yourself or contributing them to the public account. Each ECU that you keep raises your end of period income by 1 ECU. Supposing you contributed this point to the public account instead, then the total contribution to the public account would rise by 1 ECU. Your income from the public account would thus rise by $0.6 \times 1 = 0.6$ ECU if you are a Type-A member and $0.3 \times 1 = 0.3$ ECU if you are a Type-B member. However, the income of the other group members would also rise by 0.6 or 0.3 ECUs each, so that the total income of the group from the public account would be increased by $2 \times 0.6 + 2 \times 0.3 = 1.8$ points. Your contribution to the public account therefore also raises the income of the other group members. On the other hand you also earn an income for each point contributed by the other members to the public account. In particular, for each point contributed by any member you earn either 0.3 or 0.6 ECUs depending on your type.

The second stage (Allocation Stage 1)

In the second stage you will be informed about how much each group member contributed individually to the public account in the first stage. In this stage you can **reduce or leave equal the earnings of each member of your group by distributing points**. The other group members can also reduce your income if they wish to.

To reduce another player's earnings you will have to distribute points. The fee for assigning points equals 1 ECU irrespective of (a) the number of points you choose to assign and (b) the number of players to which you assign points. That is, you will pay 1 ECU whether you assign 1, 2, 3 or more points to a single player or 1, 2, 3 or more points to two or three players. To assign points you will use a screen as the one seen in Figure 3.

Every point you assign to another player reduces their earnings by 1 ECU. Similarly, your earnings will be reduced by 1 ECU for every point you receive from your group members.

You may distribute as many points as you wish to a given player. However, the total number of points you assign to a given player cannot exceed that player's earnings from the contribution stage.

Example 1: Suppose that you give 2 points to player 1. This costs you 1 ECU and reduces player 1's income by 2 ECU.

Example 2: Suppose that you give 4 points to player 1 and 3 points to player 2. This costs you 1 ECU and reduces player 1's earnings by 4 ECU and player 2's earnings by 3 ECU.

Period

1

Remaining time [sec]: 42

Allocation Stage 1

Please use the fields below to assign points to the other players. If you don't wish to assign points to a particular player you must enter '0'.

Player	Return from public account	Contribution	Earnings from contribution stage	Points you assign
Player 1	0.60	15	21.80	
You	0.60	5	31.80	
Player 3	0.30	8	20.40	
Player 4	0.30	0	28.40	

OK

Figure 3 – Allocation stage 1

Your total earnings from the two stages are therefore calculated as follows:

Total earnings (in ECUs) at the end of the second stage (allocation stage 1)

= Earnings from the 1st stage – Points you receive (– 1 if you assign points)

If the number of points that you receive *across* stages exceeds your first stage earnings, participants cannot assign any more points to you. In addition, **all points exceeding your earnings from the 1st stage will not be counted in determining your earnings from the stage.** The following example illustrates this point.

Example 2: Suppose that your earnings at the end of the 1st stage are 10.5 ECU and you are assigned 12 points in total. If you have not assigned points to others, your earnings will be $10.5 - 10.5 = 0$ ECU.

Recall that the fee for assigning points equals 1 ECU. Therefore, your earnings in ECU after the second stage can be negative. The lowest possible ‘earnings’ you can have from a period is -1 ECU. If your earnings are negative at the end of the stage, this will be covered by the 111 ECU that we gave you in the beginning of the experiment (show-up fee).

If none of the members of your group distributes points then the period finishes and the next period begins again with stage one. Otherwise, a third stage will follow.

The third stage (Allocation Stage 2)

In the third stage, you will be informed of the points that each person in your group assigned to you and the other members in your group. Similarly, the other members of your group will be informed about how many points you assigned to each of them. In addition, you will be reminded of the earnings each group member had after the contribution stage, and the total points each group member has received in total up to this point. Then you can again **reduce or leave equal the earnings of each member of your group by distributing points**. As in Allocation Stage 1, other group members can also reduce your earnings if they wish to. To assign points you have to use a screen similar to the one seen in Figure 4.

Note that in this stage **you do not have to pay the fee to reduce the earnings of others if you have already assigned points in the previous stage**. You may always assign points even if the fees would make your earnings negative.

The number of points that you can assign to a player can not exceed the earnings of a player taking into account the points that he has already been assigned. Thus, if an individual began with earnings of 20 ECU and was assigned 16 points in allocation stage 1, the maximum number of points you could assign to him/her in allocation stage 2 is 4.

Note that even if your earnings become zero (as a result of being assigned points by others), you will always be able to assign points as long as some individuals have positive earnings.

Example 3: Suppose your earnings at the end of stage 1 were 20 ECU. You chose to assign points to player 1 in stage 1. You also chose to assign points to player 1 and 2 in stage 2. No points were assigned to you. Therefore, your earnings will equal $20 - 1 = 19$ ECU.

Example 4: Assume you are player 2 in Figure 4. Your earnings after the contribution stage were 31.80 ECU. As you can see in Figure 4, in Allocation Stage 1 players 1 and 4 each assigned 2 points to you. Therefore, your earnings are reduced by 4 ECU. You also assigned 3 points to player 4 which costs you 1 ECU. Therefore, your earnings at the start of Allocation Stage 2 will be $31.80 - 4 - 1 = 26.80$ ECU.

Period
1
Remaining time [sec]: 49

Allocation Stage 2

In the table below you can see the points that were assigned to each player in your group during Allocation Stage 1:

Player	Return from public account	Contribution	Earnings from contribution stage	Total points received	Points assigned by Player 1	Points assigned by YOU	Points assigned by Player 3	Points assigned by Player 4
Player 1	0.60	15	21.80	0		0	0	0
You	0.60	5	31.80	4	2		0	2
Player 3	0.30	8	20.40	0	0	0		0
Player 4	0.30	0	28.40	3	0	3	0	

Please use the fields below to assign points to the other players. If you don't wish to assign points to a particular player you must enter '0'.

Points you assign to Player 1

Points you assign to Player 3

Points you assign to Player 4

OK

Figure 4 – Allocation stage 2

If none of the members of your group distributes points then the period finishes and the next period begins again with stage one. Otherwise, a fourth stage will follow.

Fourth stage (Allocation stage 3) and beyond

Your **task** in the fourth stage and beyond is the **same as in stage 3**. After being informed of the points distributed in your group you will be able to assign further points. The costs of assigning points, as well as the earnings reduction caused by each point remain the same as before. That is, if you have paid the fee to assign points to any player, you will not have to incur a cost to assign further points to any player. As in previous periods, the implications of assigning or receiving points, as well as the restriction on the number of points that can be assigned to each player remain the same as before.

When does a period end?

A period ends and a new one begins when one of the following occurs.

- No points are distributed in a given stage.
- Points are distributed, but no player would be allowed to assign any more points if another stage followed. This can happen if the points assigned to all players in the group are equal or greater than their earnings from the first stage.

Below is an example illustrating when a period ends. (As all examples in the instructions, the entries should not be taken as a guide for behavior in the experiment.)

Example:

Assume that after the contribution stage, the payoffs are as follows:

Player 1: 20 ECU

Player 2: 25 ECU

Player 3: 30 ECU

Player 4: 35 ECU

Assume that players have assigned points in Allocation Stage 1, 2 and 3. Further, assume that after Allocation Stage 3 the total number of points allocated to each player is:

Player 1: 16 points

Player 2: 27 points

Player 3: 30 points

Player 4: 32 points

As the number of points assigned to players 2 and 3 is greater or equal to their earnings from the first stage, no further points can be assigned to them. In addition, only the first 25 points will be counted in determining player 2's earnings.

Assume that in Allocation Stage 4 player 4 assigns 4 points to player 1 and no other player assigns points. Hence the total number of points allocated to each player is:

Player 1: 20 points

Player 2: 27 points

Player 3: 30 points

Player 4: 32 points

As points were assigned, Allocation Stage 5 will be entered. Notice, however, that now points can only be assigned to player 4, the maximum number of which is 3 as otherwise the total number of points would exceed player 4's earnings from the contribution stage.

If no points are assigned at Allocation Stage 5 or if 3 points are assigned to player 4 the period will end. At the end of the period you will receive a summary of what happened in the period. The format of this summary can be seen in Figure 5.

The period ends because no player assigned any points in the last stage.	
Other groups are still making decisions in this period. While you wait for the experiment to continue, please take the next few minutes to complete a summary sheet for this period. These notes may help you make decisions in future periods.	
Your contribution:	5
Contribution of Player1:	15
Contribution of Player3:	8
Contribution of Player4:	0
Total points assigned to you:	4
Total points assigned to Player1:	0
Total points assigned to Player3:	0
Total points assigned to Player4:	3
Your earnings from this period are:	26.80

Figure 5 – Summary screen at the end of a period

A paper summary sheet is available for each period on which you can keep track of any events that occurred in a period. These notes will help you make decisions in future periods. If you have any further questions please raise your hand and one of the supervisors will come to help you. Otherwise please answer the control questions.

Control Questionnaire (see formulas on page 7 and 10)

1. Each group member has an endowment of 20 ECUs. Nobody (including you) contributes any ECUs to the public account. How high are:
 - a. The earnings of Type-A players after the first stage?
 - b. The earnings of Type-B players after the first stage?
2. Each group member has an endowment of 20 ECUs. You contribute 20 ECUs to the public account. All other group members contribute 20 ECUs each to the public account. Suppose that you are a player of Type-A
 - a. Your earnings after the first stage?
 - b. The earnings of the other Type-A group members?.....
 - c. The earnings of the Type-B group members?
3. Each group member has an endowment of 20 ECUs. Suppose that you are a player of Type-A and you contribute 20 ECUs to the public account. The other Type-A also contributes 20

ECUs, whereas the two Type-B player each contribute 5 ECUs (10 ECUs together). How high are:

- a. Your earnings after the first stage?
 - b. The earnings of the other Type-A group members?.....
 - c. The earnings of the Type-B group members?
4. Each group member has an endowment of 20 ECUs. Suppose that you are a player of Type-A and you contribute 20 ECUs to the public account. The other Type-A also contributes 20 ECUs, whereas the two Type-B player each contribute 10 ECUs (20 ECUs together). How high are:
- a. Your earnings after the first stage?
 - b. The earnings of the other Type-A group members?.....
 - c. The earnings of the Type-B group members?
5. Each group member has an endowment of 20 ECUs. The other three group members contribute together a total of 30 ECUs to the public account. Suppose that you are a player of Type-A. What is:
- a. Your earnings after the first stage if you contribute 0 ECUs to the public account?
 - b. Your earnings at the end of the period if you contribute 20 ECUs to the public account?
6. Each group member has an endowment of 20 ECUs. You contribute 8 ECUs to the public account. Suppose that you are a player of Type-B. What is:
- a. Your earnings after the first stage if the other group members together contribute a further total of 12 ECUs to the public account?.....
 - b. Your earnings at the end of the period if the other group members together contribute a further total of 22 ECUs to the public account?.....
7. Your earnings from the first period are 25 ECU. How much will your earnings at the end of the second stage (allocation stage 1) be if:
- a. You receive 2 points, but do not assign any yourself?
 - b. You receive 2 points and assign 3 points yourself to a single group member?.....
8. Assume you assign 2 points to another group member, no one else in your group assigns any points and all members in your group have a positive payoff. Will another stage follow?
9. Assume no member of your group assigns points including you. Will another stage follow?
10. Assume the earnings of Player 2 after the contribution stage are 25 ECU. Assume also that Player 1 assigns 25 points to Player 2. Will Player 2 be able to assign further points in this period?

Instructions – Part 3

This part of the experiment will last 10 periods (periods 11-20). Recall that at the beginning of the experiment, you were randomly divided into groups of four. **The composition of your groups for this part of the experiment is the same as in Part 1 and as in Part 2 and will remain the same throughout the remainder of the experiment.** Additionally, the identification numbers of Part 2 and your type (“Type A” or “Type B”) will also remain valid in this part.

Moreover, for the ten periods of this part, you will face exactly the same decision problem and the same stages as in part 2. **The only difference is that you will be able to communicate with other members in your group in periods 11, 14, and 17.**

Details about Communication (Periods 11, 14, 17):

At the start of periods 11, 14, and 17, you will have the opportunity to communicate with other members in your group. Communication in these periods will take the form of text messages in two “chat boxes”. You can use a chat box in which your messages are sent to all other group members or you can use a chat box in which your messages are only sent to the other participant of the same type as you (Type A or Type B). These messages will not be viewed by the other two group members of the other type. When you type a message, your identification number (e.g. Player 1) will appear before the message as well as you type.

Communication Rules

During a communication period, you can discuss anything you like, including what you think is the best approach to the experiment, what you plan to do in the following period, what happened in previous periods, or what you would like others to do. However, there are two restrictions on the types of messages that you may send. First, you may not send a message that attempts to identify you to other group members other than by your identification number. Thus, you may not use your real name, nicknames, or self-descriptions of any kind (“Tom Smith here,” “I’m the guy in the red shirt sitting near the window,” “It’s me, Sandy, from French class,” or even “As a woman [Latino, Asian- American, etc.], I think. . .”). The second restriction is that there must be no threats or promises pertaining to anything that is to occur *after the experiment ends*. We will not check communication messages during the experiment. But if a participant complains about a rule violation and if the complaint is valid, the responsible subject will be penalized: This person will not receive any earnings from the experiment other than the show-up fee.

Period 11	Remaining time [sec]: 169
--------------	---------------------------

Before you decide how much you would like to contribute to the public account in the next period, you are now able to send messages to the other participants.

Use the box at the top to send messages to all participants, irrespective of their type.
Use the box at the bottom to send messages to the other participant with the same type as you.

Messages to all other participants:

Player 1 (Type B): Test 1
Player 2 (Type B): Test 2
Player 3 (Type A): Test 3
Player 4 (Type A): Test 4

Messages to the participant with the same type as you:

Player 1 (Type B): Test a
Player 2 (Type B): Test b

Figure 1 – Communication Screen

If you have any further questions please raise your hand and one of the supervisors will come to help you.

Appendix G – Instructions for communication coders

This appendix contains the instructions that were given to the three independent coders to classify our communication data.

Instructions

Classification of Communication Data

In the following, we will describe the classification process for our experimental data. We assume that you are familiar with public good games. We will first outline our experiment in more detail so that you understand in which environment our communication data was generated. Afterwards, we will explain in detail according to which criteria we ask you to classify the provided scripts of group communication data. Please read these instructions as well as the sample instructions of the original experiment that you will find attached carefully. After reading the instructions, we would advice you to start by classifying 3-4 groups (overall 17-18) per treatment, ask potential clarification questions and then proceed to the remaining groups.

Experiment

Overview. Our experiment consisted of three different treatments. The general structure of all of these treatments was the following: In **part 0**, subjects first did a **real-effort task**. In **part 1 and 2**, they played a **public good game** for overall 20 periods (**10 periods in each part**). Subjects' relative performance in the real-effort task determined whether they received a high or low return from the public account in part 1 and 2 of the experiment. In the different treatments, we varied whether subjects had the ability to **communicate** and to **punish** other group members, as outlined in more detail below.

The basic game. We implement a public good game. We use a simple voluntary contribution mechanism, with a linear production technology for the public good. At the beginning of each period, each of **four** participants is given an **endowment of 20** experimental currency units (ECU). Subjects have to allocate this endowment between a “private account” and a “public account”. The earnings of group member i at the end of the first stage are given by:

$$\pi_i^1 = 20 - c_i + m_i \sum_{j=1}^4 c_j$$

where m_i denotes the **return** to total allocations by the group towards the public account. There are two *high types* (so-called **Type A**) that receive a high return from the public account, $m_i = 0.6$, and two *low types* (so-called **Type B**) that receive a low return from the public account, $m_i = 0.3$. The two players **relatively performing better** in the real-effort task become the high types, receiving the higher return from the public good.

In this environment, subjects' decision variable is c_i , where $c_i \in \{0, 1, \dots, 20\}$. Crucially the game represents a **social dilemma** since the individual returns from the public good, m_i , are lower than the returns from the private account, 1, making zero contribution to the public good the dominant Nash equilibrium. Since $n \cdot m_i > 1$, total returns, however, are greater than private returns, implying that it is socially optimal to contribute the **full endowment** of 20 for all group members.

Contribution rules. Thus, although it may be **individually rational** to contribute 0, group members could still generally agree to contribute something since this could make **all group members better**. But even in case group members generally agree to contribute something, they may still **disagree about who should contribute what**. In particular, different types (A: high vs. B: low) may favour different contribution rules:

1. *Efficiency rule:* All group member (4 out of 4) contribute **20**. This leads to efficiency in the sense that group earnings are highest. Crucially, while high types earn 48 ($=20-20 + 0.6 \cdot 80$) low types only earn 24 ($=20-20 + 0.3 \cdot 80$). Thus, this rule may be favoured by type As.
2. *Equality rule:* The two high types contribute **20** and the two low types contribute **5**. This leads to equal earnings of both types since both high types ($=20-20 + 0.6 \cdot 50$) and low types ($=20-5 + 0.3 \cdot 50$) earn 30. Thus, this rule may be favoured by type Bs because it gives them a higher earning compared to the efficiency rule.
3. *Compromise rule:* The two high types contribute **20** and the two low types contribute something in **between of 5 and 20**, $20 > c > 5$ (e.g. 8, 12 or 15 but not 5 or 20). This provides a compromise between achieving high group earnings (efficiency – 1.) and reach equal earnings (equality – 2.).
4. *Other rule:* Occasionally, subjects may also agree on rules other than the three above. This could include e.g. that (a) everyone contributes 15 or (b) that high types contribute 10 and low types contribute 5. These rules do not reach the Pareto frontier as everyone could be made better off. In case of (a) everyone could contribute 20 or in the case of (b), high types could contribute 20 and low types 10.

Of course, some groups may not agree or even discuss a specific rule at all.

Treatments. In this basic framework, we implement three different treatments that vary whether subjects can **communicate** and/or can **punish** or not.

1. **Treatment 1** – In part 1, subjects can punish and communicate; in part 2, they can only punish
2. **Treatment 2** – In part 1, subjects can only punish; in part 2, subjects can communicate and punish
3. **Treatment 3** – In both parts, subjects can only communicate but not punish.

The general idea is that in case subjects have a *punishment* opportunity, all four subjects first decide how much to contribute to the public account as before. Then subjects learn how much the other group members have contributed. Afterwards, they now have the opportunity to **reduce other players' earnings by assigning (punishment) points** to them. These points are only associated with a very small cost for the person assigning them. After the first punishment stage, subjects are informed about who punished whom and the effect punishment had on everyone's earnings. Then, there is the **opportunity to counter-punish** (in case there has been initial punishment). In case counter-punishment is applied, subjects can **counter-counter punish** and so on.

Communication is implemented in the following way. Subjects can use anonymous chat boxes to discuss how to proceed in the experiment (for a fixed amount of time). An important feature of our

implementation is that there are **two different chat boxes**. One, in which subjects can communicate with the whole group and one, in which they can exchange messages only with the other player of the same type. Crucially, this chat communication is implemented only at the beginning of **period 1, 4, and 7** in case communication is implemented in **part 1** (as e.g. in Treatment 1) of the experiment and at the beginning of **period 11, 14, and 17** in case communication is implemented in **part 2** (as e.g. in Treatment 2) of the experiment. It is your task to classify this chatbox communication according to criteria outlined below.

Classification of Communication Data

You will be provided with the communication script of every group (in every treatment) of our experiment (File: *Treatment 1/2/3 Communication Data*). Depending on the treatment, each group had the chance to communicate either three (Treatment 1 and 2) or six times (Treatment 3). Broadly speaking, your task will be to classify for every group (and every communication period) ...

- 1) ... which contribution rules subjects discussed,
- 2) ... whether groups reached an agreement and
- 3) ... to what extent (and how) the same-type chatbox was used.

You will use File *Classification* to make your input.

IMPORTANT: *When classifying the communication data, please limit yourself to making inferences only from what can clearly be derived from the message stated, i.e. do not try to think about what the player might have thought.*

Details of the classification

1. Discussion of Rules

The central question here is what rules groups discussed in the communication period. More precisely, you should classify for each of the following rules whether it has been mentioned (“1”) or not (“0”) (for each communication period separately). Crucially, it suffices that a contribution rule has been mentioned. In other words, it is not necessary that groups also agree on this rule.

- a. *Efficiency {0,1}*: Group members could mention this rule e.g. by saying something like “Everyone should contribute everything”, “We should all contribute 20”, “Both type A and type B player should contribute 20.” Sometimes, the efficiency concern may even be mentioned explicitly (although this not a necessary for a positive classification): “We should all contribute 20, because it maximizes group earnings”. Finally, subjects could also refer to **control question or scenario 2** of the instructions to mention this contribution rule.
- b. *Equality {0,1}*: Group members could mention this rule e.g. by saying “Type A players should contribute 20 whereas type B players should contribute 5.”; “As 20, Bs 5”, “A should fully contribute while B should only contribute 5.” Sometimes, the equality concern may even be mentioned explicitly (although this not a necessary for a positive classification): “We should all get the same earnings. Thus, A should contribute 20 and B 5.”. Finally, subjects could refer to **control question or scenario 3** of the instructions.
- c. *Compromise {0,1}*: Group members could mention this rule (or more precisely **these rules**) e.g. by saying “Type A players should contribute 20 and type B players

should contribute 8”, “A 20 and B 15”, “As should fully contribute but B should give less than 20, e.g. 10”. Subjects may actively advertise this rule by arguing that it presents a compromise: “Let’s make a compromise. A contributes 20 and B 12”. (Crucially, you should only classify that groups discuss a *compromise* rule in case the player’s suggestion says that A contribute their full endowment or in other words 20. If this is not the case, you may consider the next rule.)

- d. *Other rules {0,1}*: Crucially, in our understanding, subjects only discuss *other rules* in case a *specific* contribution rule is suggested (**that is not one of the above**). This would include statements as “Let’s all contribute 15”, “As should contribute 15 and Bs 5”, “A 12, B 8”. Moreover, a statement like “A should at least contribute 15 and B at least contribute 8” would also fall into this category. Unspecific statements such as: “Let’s all contribute something” or “As contribute 15 and Bs contribute something” (without specifying more precisely what B does), should not be classified.

It often happens that more than one rule is discussed. In case that none of these rules are, however, discussed, please insert “0” four times.

In *non-initial* communication periods (periods 4, 7 and 14, 17), it fairly frequently happens that groups do not discuss any rules but just stick to a rule they followed in the previous periods before communicating again. Sometimes, this happens explicitly by saying “Let’s stick to what we have done so far”. Sometimes, it happens implicitly; groups e.g. do not talk at all or talk e.g. about the weather.

For the case that no explicit statement is made, we provide you with the information whether a group followed one of the four rules in the periods before and after the communication period. In case a group consistently follows a rule and does e.g. not talk at all, you should also classify this group as sticking to a rule. In both – implicit as well as explicit – cases, you should insert “1” in the “stick” column.

Generally, we would ask you to focus on the communication data and only use the data what groups actually did with great caution. Sometimes, this information maybe misleading since minor deviations from a rule can already lead to a group being classified as not having followed a rule.

Finally, we would like you to classify whether any group member makes a statement (“1”) that no (punishment) points **should not be distributed** at least once or whether no such statement is made (“0”) (“Please do not assign points!”, “No points should be assigned”)

2. Agreement

The central question here is whether a group agrees on one of the four rules discussed above. As outlined below, our definition of an agreement is **fairly strict** and we would like you to analyse the communication data very carefully whether the conditions are really fulfilled:

- *Agreement {1}*: Groups may discuss many different rules. At some point, however, one player suggests one of the contribution rules outlined above. Crucially, an agreement in our definition, then requires that **all other three player** actively agree to this rule (by saying e.g. “OK”, “Let’s do this”, “☺”, etc.). Thus, only when all four group members actively favour one of the discussed rules an agreement is reached.

Moreover, in case such an agreement has been reached but afterwards a further suggestion of a contribution rule is made, this suggestion has either to be withdrawn from this person (e.g. “Ok, let’s follow your suggestion”) or the other three players have to actively agree with the new suggestion.

We only relax this strict definition for **non-initial communication** periods. As outlined before, groups sometimes **explicitly (or implicitly) stick to a rule** without any further discussion (or renewed agreement) in later communication periods. You should also classify this group as having reached an agreement. Thus, in this instance, it is **not** necessary that all four people actively agree again on a rule (or even actively agree to stick to a previous rule) as long as no (new) rules are discussed explicitly or active discontent with a previous rule is mentioned. In case, (new) rules are discussed, all group members have to actively agree again to this rule for agreement classification as outlined before.

- *(No) Agreement {0}*: This classification includes many possible scenarios. A player may suggest a contribution rule but only 2 out of the 3 other players actively agree to this rule. Sometimes, the third player may just not agree because the time is already over. It may also happen that two different contribution rules are discussed and players just do not reach an agreement which of those two rules to follow. Moreover, it could happen that subjects discuss the weather or only agree on unspecific contribution rules (“Let’s all contribute something”).

In case, you classify a group as not having reached an agreement {0}, we would like you to rank the *level of hostility* that those groups experience. This ranking should be from:

{1} *very friendly atmosphere* to {5} *very hostile atmosphere*⁷³

An example of a *1-ranking* might be a group in which the fourth person just fails to agree to a contribution rule because time has run out. An example of a *3-ranking* might be group that discusses alternative rules but cannot agree on one rule. Crucially, players remain polite during the whole discussion. An example of *5-ranking* might be a group that cannot agree on a rule and in which one player (or one type) **threatens** to punish the other player by assigning points in case they do not follow his or her suggestion (“We will all leave with nothing in case you do not contribute X.”, “I will assign points to you as long as you do not contribute X”) **or** in which subjects use **insulting language** (“Equality you douche”, “You are an asshole, player 1!”) and have difficulties to return to a more civilized discussion.⁷⁴ In case, you classify the hostility level in a group with a 4 or 5, we ask you to comment what incidences made you choose this classification (e.g. “*threats*”, “*insults*”, or something else).

3. *Same-Type chatbox:*

The central question here is whether group members use the same-type (second) chatbox and what they use it for.

⁷³ In the main text, for convenience, we describe the scale as running from 0 to 4.

⁷⁴ In treatment 3 (*ComOnly*), subjects cannot threaten to punish other subjects by assigning points since no punishment is implemented. Participants may, however, still “threaten” to contribute nothing in case others do not follow their suggestions. Depending on your impression, we would, however, classifying this kind of threat only with a 3 or 4 since it potentially conveys less hostility.

First, we would like you to classify whether the **same-type** chatbox was used {1} or not {0}:

- *Chatbox used {1}*: Crucially, you should only classify that the chatbox was used in case group members **discuss how to proceed in the game**, e.g. “We should persuade the other to contribute more”, “Wouldn’t it be best in case we contribute 20?”, “What should we do?” etc.
- *Chatbox not used {0}*: In case no entry at all is made. But also in case subjects use the same-type chatbox but do not discuss how to proceed in the game. Subjects may e.g. just say “Hi” without any further discussion about how to proceed or talk about the weather or other things.

Second, in case you classify that the same-type chatbox has been used, we would like you to state whether it has been used to discuss to specific issues (beyond a general discussion about how to proceed). Namely, whether it has been used discussing (not necessarily agreeing) whether and/or under which circumstances **the other types should be punished** {1} (“Let’s punish the others, in case they do not stick to our agreement”, “Let’s punish the other in case they contribute less 20”). Additionally, we would like you to state whether two players discuss in the same-type chatbox (not necessarily agree) **to deviate from suggestions or agreements made in the all-player chatbox** {2} (“In the last round, let’s contribute nothing!”, “Let’s contribute 20 in first round, but afterwards, we should contribute nothing”). In case, it has been used for both aspects, please insert “1,2”.

Please start by classifying 3-4 of each treatment. After asking clarification questions – in case you have any – please proceed to classifying the remaining groups.

Appendix H – References

- Fehr, E. and K. Schmidt (1999). A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics* 114(3), 817-868.
- Gneezy, U., A. Kajackaite, J. Sobel (2018). Lying aversion and the size of the lie. *American Economic Review* 108 (2), 419-453.
- Lundquist, T., T. Ellingsen, E. Gribbe, and M. Johannesson (2009). The aversion to lying. *Journal of Economic Behavior & Organization* 70 (1-2), 81-92.