

Stenzel, Aurel; Lohse, Johannes; Requate, Till; Waichman, Israel

Working Paper

Altruistic Cooperation

CESifo Working Paper, No. 11880

Provided in Cooperation with:

Ifo Institute – Leibniz Institute for Economic Research at the University of Munich

Suggested Citation: Stenzel, Aurel; Lohse, Johannes; Requate, Till; Waichman, Israel (2025) : Altruistic Cooperation, CESifo Working Paper, No. 11880, CESifo GmbH, Munich

This Version is available at:

<https://hdl.handle.net/10419/320101>

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

Altruistic Cooperation

Aurel Stenzel, Johannes Lohse, Till Requate, Israel Waichman

Impressum:

CESifo Working Papers

ISSN 2364-1428 (electronic version)

Publisher and distributor: Munich Society for the Promotion of Economic Research - CESifo GmbH

The international platform of Ludwigs-Maximilians University's Center for Economic Studies and the ifo Institute

Poschingerstr. 5, 81679 Munich, Germany

Telephone +49 (0)89 2180-2740, Telefax +49 (0)89 2180-17845, email office@cesifo.de

Editor: Clemens Fuest

<https://www.ifo.de/en/cesifo/publications/cesifo-working-papers>

An electronic version of the paper may be downloaded

· from the SSRN website: www.SSRN.com

· from the RePEc website: www.RePEc.org

· from the CESifo website: <https://www.ifo.de/en/cesifo/publications/cesifo-working-papers>

Altruistic Cooperation

Aurel Stenzel¹, Johannes Lohse², Till Requate¹, and Israel
Waichman³

¹University of Kiel, Germany, `stenzel@economics.uni-kiel.de`,
`requate@economics.uni-kiel.de`.

²Leuphana University Luneburg, Germany, and University of Birmingham, United
Kingdom,
`johannes.lohse@leuphana.de`

³Bard College Berlin, Platanenstrasse 24, 13156 Berlin, Germany,
`i.waichman@berlin.bard.edu`

April 27, 2025

Abstract

We characterize ‘Games of Altruistic Cooperation’ as a class of games where cooperation leaves the individual and the group of decision-makers worse off than defection, but favours individuals outside the group. An example is climate change mitigation. In this context, we experimentally investigate whether decentralized institutions using costly punishment and/or communication support altruistic cooperation to sustain the welfare of future generations. Without punishment or communication, cooperation is low; communication alone even increases the incidence of zero contributions. However, combining peer punishment with communication strongly increases cooperation, showing that an effective decentralized solution to a Game of Altruistic Cooperation exists.

Keywords: Games of altruistic cooperation; social dilemma; intergenerational good game; punishment ; behavioral experiment

JEL classification code: C92; D74; H41; Q54

1 Introduction

Collective action problems arise whenever “an individual can benefit from mutual cooperation, [but] each one can also do even better by exploiting the cooperative efforts of others” (Axelrod, 1981, p.1391). Such conflicts between private and public interests contribute to inefficiencies in various economic interactions (e.g., Olson, 1965; Hirshleifer, 1978; Dawes, 1980; Bergstrom et al., 1986; Dawes and Thaler, 1988; Gardner et al., 1990; Ostrom et al., 1992; Kosfeld and Rustagi, 2015; Bonatti and Hörner, 2011; Chan, 2016; Becker, 2015). While non-cooperative game theory predicts that individuals will refrain from cooperating, experimental research has consistently shown that cooperation occurs far more frequently than the theory anticipates. Yet, in repeated game settings, maintaining high levels of cooperation over time remains a persistent challenge (Ledyard, 1994; Zelmer, 2003; Villeval, 2020). Nonetheless, other experimental studies suggest that institutional mechanisms—particularly sanctioning (e.g. Fehr and Gächter, 2000; Andreoni et al., 2003) and communication (e.g. Isaac and Walker, 1988a; Balliet, 2010)—can substantially foster cooperative behavior.

Much of this experimental evidence stems from classical social dilemmas, such as the prisoners’ dilemma (Schelling, 1973; Axelrod, 1981), the public good game (Dawes and Thaler, 1988; Isaac and Walker, 1988b; Fehr and Gächter, 2000), and the common pool resource game (Hardin, 1968; Ostrom et al., 1992, 1994). These dilemmas share a common incentive structure where (a) defection is the best response when all other decision makers cooperate, and (b) mutual cooperation leads to a *higher* payoff than mutual defection for all decision makers (Dawes, 1980; Dawes and Thaler, 1988; Beckenkamp, 2006).¹ Thus, if decision makers were to maximize their payoff as a group or to follow a social planner who maximizes group payoffs, they would cooperate.

Classical social dilemmas therefore capture the tension between cooperation and defection within a well-defined group of decision makers who mutually benefit from each other’s cooperation. In large, interconnected economies, however, cooperation problems often extend beyond cases where decision makers and beneficiaries are the same. This occurs when the benefits of cooperation accrue to others, either spatially, temporally, or both. Take climate change as an example, where current decision makers decide whether to cooperate, i.e., transfer some of their wealth to future generations. This cooperation problem appears particularly difficult to resolve: despite potentially large long-term benefits from reducing carbon emissions, individual efforts fall short of meeting the 2°C Paris target (Roelfsema et al., 2020). One reason why cooperation fails in this context is that the benefits extend beyond the immediate group of decision makers. Since decision makers do not directly benefit from their own mitigation efforts, the logic of cooperation in these situations differs from that in classical social dilemmas. In particular, cooperation now requires decision makers to

¹In some classical social dilemmas, such as the prisoner’s dilemma and the linear public good game, defection is a dominant strategy. Condition (a) is more general as it applies also to other social dilemmas, such as the common pool resource game (e.g., Ostrom et al., 1992).

sacrifice some of their own current wealth for the sole benefit of future others. More generally, we define this class of cooperation problems as “games of altruistic cooperation”. In this class of games, condition (a) of the classical social dilemma still holds (i.e., defection is a best response if everyone else cooperates). But condition (b) is replaced by the following alternative condition (b*): mutual cooperation leads to a *lower* payoff for each decision maker. Thus, even if decision makers were acting as a cartel optimizing their group payoff, they would collectively choose to deviate from the globally optimal allocation that an intertemporal social planner would choose.

Games of Altruistic Cooperation hence pose a greater challenge to cooperation because there are fewer motives to cooperate than in classical social dilemma games. In classical social dilemmas, cooperation may arise from pure altruism, reciprocity, or strategic considerations (Axelrod, 1981; Kreps et al., 1982; Rand and Nowak, 2013). By contrast, in Games of Altruistic Cooperation, successful cooperation cannot be due to direct reciprocity or strategic motives. This raises new questions regarding the institutions that can support altruistic cooperation, which we address in this study.

A prominent example of a Game of Altruistic Cooperation is the Intergenerational Goods Game (IGG, Hauser et al., 2014). In this game, a group of decision makers (hereafter referred to as “generation”) can contribute to a public good. Different generations form a sequence in which later generations benefit from the contributions of earlier generations. If the total contributions of the current generation reach a predefined threshold, the next generation receives a new endowment and can participate; if not, the game ends with the current generation. This structure implies that the provision of the public good is not in the monetary interest of either the individual decision maker or the current generation as a whole. However, overall efficiency across generations is maximized when each generation contributes sufficiently to reach the threshold, resulting in sustained cooperation over time.

Previous evidence, notably Hauser et al. (2014), Camerer et al. (2018), and Lohse and Waichman (2020), suggests that while the majority of participants in IGG experiments contribute to a public good that only benefits later generations, their total contributions often fall short of reaching the threshold (resulting in low levels of “sustained generations”). Moreover, Hauser et al. (2014) and Balmford et al. (2024) demonstrate that median voting is an effective institution that helps most generations to reach the threshold. However, implementing the outcome of the vote requires centralized enforcement, which is often absent in international settings such as in climate change negotiations where votes are non-binding. Without centralized enforcement, where votes are only partially binding, the effectiveness of voting is significantly reduced (Hauser et al., 2014). A fully decentralized institution that has been studied is costly peer punishment. In contrast to findings from classical social dilemmas, where peer punishment is typically highly effective, it only moderately increases successful cooperation in the IGG (Lohse and Waichman, 2020). Overall, there is no evidence that decentralized institutions can achieve high levels of altruistic cooperation in the IGG.

After providing a formal definition of games of altruistic cooperation, we present the results of an IGG experiment that investigates the potential of decentralized institutions to support altruistic cooperation. Specifically, we examine whether combining peer punishment with communication – both of which could be considered as naturally-evolved institutions for solving cooperation problems within a fixed group (Smith, 2010; Markussen et al., 2014; Kamei et al., 2015) – results in a high level of sustained generations. We set up four treatments: a baseline with no communication or punishment, a treatment in which members of the current generation can punish peers at a cost, a treatment in which members can communicate, and a treatment that combines both punishment and communication.

Both punishment and communication have been extensively studied in the context of classical social dilemmas. Peer punishment has been found to be effective in facilitating cooperation in linear public good games (Fehr and Gächter, 2000, 2002). Communication typically enhances cooperation in public good games (Isaac and Walker, 1988a; Bochet et al., 2006; Bochet and Putterman, 2009) and in threshold public good games (Bicchieri, 2002; Tavoni et al., 2011; Palfrey et al., 2017). The joint use of punishment and communication is particularly effective in classical social dilemmas where each institution alone has a small effect (Ostrom et al., 1992; Janssen et al., 2010; Cason and Gangadharan, 2016).² However, because the incentive structure in Games of Altruistic Cooperation is fundamentally different from that in classical social dilemmas, it is unclear whether these institutions will be equally effective in facilitating altruistic cooperation.

Our main findings are as follows: In the absence of any additional institution, only a small fraction of generations (20%) achieve sufficient cooperation to reach the threshold. Punishment alone more than doubles the likelihood of reaching the threshold, but the share of sustained generations is still only 48%. Communication alone does not significantly increase successful cooperation (27%) and even increases the incidents of zero contributions. However, when punishment and communication are combined, the number of sustained generations increases significantly, approaching the maximum number possible (87%). This large increase is surprising because neither of the two institutions affects the basic incentive structure fundamentally: mutual defection still maximizes the joint payoff for the current generation (as well as for each individual). We also identify a potential channel leading to this outcome. With peer punishment, the nature of communication changes: decision makers now rarely suggest defection, and the content of the messages exchanged tends to focus on cooperation instead. In general, the expression of cooperative intentions, together with enhanced coordination through communication, makes a combination of punishment and communication highly successful in inducing altruistic cooperation.

²This finding typically arises when the payoff from cooperation is not constant, leading to cooperative outcomes that are more difficult to coordinate on because some intermediate level of cooperation is optimal.

2 Classical social dilemmas and Games of Altruistic Cooperation

Consider a strategic form game with the set of players $N = \{1, \dots, n\}$. Within this set, there is a subset of decision makers $D = \{1, \dots, m\} \subseteq N$, with $m \leq n$, while $P = \{m+1, \dots, n\}$ is a subset of passive players. Each decision-maker $i \in D$ has a strategy set S_i . We let

$$\mathbf{S} = \times_{i \in D} S_i$$

denote the set of all possible strategy profiles for decision makers. A strategy option for decision-maker i is denoted by s_i , and the strategy profile of decision makers other than i is denoted by s_{-i} . We write $\mathbf{s} = (s_1, \dots, s_m) \in \mathbf{S}$ for a strategy profile chosen by all decision makers. The payoffs of all players (decision makers and passive players) depend only on the options chosen by the decision makers. The payoff for player $i \in N$ is denoted by $\pi_i(\mathbf{s})$, while the aggregate payoff for all players is denoted by $\pi(\mathbf{s}) = \sum_{i=1}^n \pi_i(\mathbf{s})$. Note that the aggregate payoff includes the payoffs of both the decision makers and the passive players.

2.1 Classical social dilemmas

Classical social dilemmas are characterized by the following properties. Each strategy set S_i contains at least two possible strategies: a cooperative strategy, denoted as c_i , and a defecting strategy, denoted as d_i .³ We denote the strategy profile for mutual cooperation (i.e., each decision maker i chooses c_i) as \mathbf{c} and for mutual defection (i.e., each decision maker i chooses d_i) as \mathbf{d} . Furthermore, the profile of a single decision maker i , who unilaterally deviates from cooperation and instead chooses to defect while the others continue to cooperate, is denoted by (d_i, c_{-i}) . The classical dilemma arises from the following conditions (see, e.g., Dawes, 1980; Dawes and Thaler, 1988; Beckenkamp, 2006):

Condition a: Defection is the best response for a decision maker $i \in D$ if all other decision makers cooperate; i.e., $\pi_i(d_i, c_{-i}) \geq \pi_i(c_i, c_{-i})$.

Condition b: Mutual cooperation leads to a higher or equal payoff than mutual defection for all decision makers; i.e., $\pi_i(\mathbf{c}) \geq \pi_i(\mathbf{d})$ for all $i \in D$. Furthermore, mutual cooperation leads to a higher or equal aggregate payoff for all players (i.e., decision makers and passive players) than mutual defection; i.e., $\sum_{i=1}^n \pi_i(\mathbf{c}) = \pi(\mathbf{c}) \geq \pi(\mathbf{d}) = \sum_{i=1}^n \pi_i(\mathbf{d})$.

A seminal example of a classical social dilemma is a linear public good game (e.g., Dawes and Thaler, 1988; Isaac and Walker, 1988a; Fehr and Schmidt, 1999). There are $n \geq 2$ players who simultaneously decide on their contributions

³For simplicity, we implicitly assume that the defecting strategy is unique. In more general settings, a decision maker's unilateral optimal response may, of course, depend on the overall strategy profile of all other players.

$x_i \in [0, w]$ to a public good. That is, in this particular example, the set of players is equal to the set of decision makers. We denote x_{-i} as the vector of contribution decisions of the $n - 1$ players other than i . The monetary payoff of player i is given by

$$\pi_i(x_i, x_{-i}) = w - x_i + a \sum_{j=1}^n x_j$$

with $\frac{1}{n} < a < 1$. Each player's cooperative (defecting) strategy choice is to contribute the whole endowment $c_i = w$ (nothing, i.e. $d_i = 0$). Mutual cooperation leads to the highest payoff for each individual player (and hence the highest aggregate payoff) provided that a is greater than $\frac{1}{n}$ (Condition b). However, the best response for each player is to contribute nothing since this leads to the highest individual payoff if everyone else cooperates (Condition a).⁴

2.2 Games of Altruistic Cooperation

A different type of social dilemma arises when only a subset of society is responsible for decision making and when the interests of these decision makers diverge from those of the broader population. This is exemplified by some of the major challenges we face today, such as the prevention of climate change, which requires current generations to bear the entire cost of cooperation while future generations reap all the benefits from it (Weiss, 1990). Other challenges can be spatial in nature, such as when a group of decision makers decides on a public project (such as building a dam) that later helps or harms others located in a different region or even a different country through a downstream externality (Delaney and Jacobson, 2014).

Here we describe the general properties of Games of Altruistic Cooperation, in which decision makers must sacrifice their own payoff in order to increase the aggregate payoff (of all players). Therefore, Games of Altruistic Cooperation are characterized by Condition a — as defined in the previous section — as well as Condition b*.

Condition b*: Mutual cooperation leads to a lower or equal payoff than mutual defection for all decision makers; i.e., $\pi_i(\mathbf{c}) \leq \pi_i(\mathbf{d})$ for all $i \in D$. However, as in b), mutual cooperation leads to a higher or equal aggregate payoff for all players (i.e., decision makers and passive players) than mutual defection; i.e., $\sum_{i=1}^n \pi_i(\mathbf{c}) = \pi(\mathbf{c}) \geq \pi(\mathbf{d}) = \sum_{i=1}^n \pi_i(\mathbf{d})$.

Consider as the simplest possible example of a Game of Altruistic Cooperation a modification of the standard Dictator Game (Forsythe et al., 1994), in which multiple decision-makers allocate funds from a private endowment to be multiplied (with a factor larger or equal than 1) and subsequently shared with passive recipients. This simple modified dictator game captures condition a and

⁴In a linear public good game, defection is a dominant strategy since $a < 1$, and therefore, the marginal investment in the public good causes a loss of $(1 - a)$ for the individual player.

b* of games of altruistic cooperation: first, defection (contributing nothing) is the optimal strategy for each individually rational decision maker; second, mutual cooperation (contributing the full endowment or any positive amount) leaves decision makers strictly worse off individually and as a group compared to defection (withholding contributions entirely). Yet, with a sufficiently large efficiency multiplier, full cooperation maximizes aggregate payoffs for all players (i.e., active decision-makers and passive recipients).

In his meta-analysis, [Engel \(2011\)](#) observes that 28.35% of the initial endowment is donated in standard dictator games (with single dictators and recipients and without an efficiency multiplier). This suggests that participants are willing to cooperate even when it cannot be explained by selfish motives.⁵ Furthermore, efficiency considerations matter (albeit the overall effect size in the meta analysis is small). Further evidence for altruistic preferences comes from a literature on charitable giving, which also shows that people are willing to share some resources with a charitable recipient and that changes to the price of giving through matching grants or rebates (that work similar to the factor m increase generosity, e.g., [Karlan and List, 2007](#); [Vesterlund, 2016](#)). In the type of modified multiplayer dictator game presented above, [Panchanathan et al. \(2013\)](#) finds that dictators transfer less and recipients earn less when there are multiple dictators (28%, 12%, and 9% of the respective endowments with one two and three dictators). This effect persisted despite mechanisms that eliminate or reduce uncertainty about the behavior of fellow dictators (using a strategy method and communication among dictators). Regarding their last treatment with communication, there is contradicting evidence about the effect of group decision making in dictator games. In line with [Panchanathan et al. \(2013\)](#), [Luhan et al. \(2009\)](#) find that groups are less generous than individuals. But [Cason and Mui \(1997\)](#) find the opposite.

While findings from dictator games suggest that some level of altruistic cooperation exists in simple interactions, it is unclear if it would be sufficient to sustain an intergenerational public good such as climate change mitigation. Such goods are often provided within a sequence of generations where later generations first assume the role of passive players and later on act as decision makers. Moreover, successful altruistic cooperation here hinges on reaching a threshold or target level through collective action. The Intergenerational Goods Game (IGG, [Hauser et al., 2014](#)) expands upon this core idea by introducing a chain of sequential generations and a collective contribution threshold, which, if met, ensures multiplied endowments are passed on to the next generation of decision makers.

⁵[Bardsley \(2008\)](#) and [List \(2007\)](#) show that it is problematic to infer an individual's social preferences from a specific dictator game, as factors such as framing (giving, taking), efficiency, and the action set's range may matter. Yet, there is evidence that (some) people 'give' money, even when they can 'take', thus possessing (some form of) altruistic preferences.

3 The repeated Intergenerational Good Game

The Intergenerational Good Game (IGG, [Hauser et al., 2014](#)) lends itself for further study for three reasons: (i) it involves a threshold, which effectively establishes the group as the decision-making unit, as no single decision maker can unilaterally secure the payout for subsequent players. This mirrors one key characteristic of the climate change dilemma; (ii) previous experiments indicate that, in the absence of supporting institutions, altruistic cooperation remains very difficult to achieve ([Hauser et al., 2014](#); [Camerer et al., 2018](#); [Lohse and Waichman, 2020](#)); and (iii) thus far, only a perfectly centralized institution has been shown to achieve considerable levels of altruistic cooperation ([Hauser et al., 2014](#); [Lohse and Waichman, 2020](#)).

In the repeated version of the IGG we newly introduce, a cohort of 12 players is randomly divided into four successive generations, each consisting of three players. Without loss of generality, we assume that players 1, 2, and 3 are assigned to Generation 1, players 4, 5, and 6 to Generation 2, players 7, 8, and 9 to Generation 3, and players 10, 11, and 12 to Generation 4. Accordingly, we define $G_1 = \{1, 2, 3\}$, $G_2 = \{4, 5, 6\}$, $G_3 = \{7, 8, 9\}$, and $G_4 = \{10, 11, 12\}$. The IGG starts with Generation 1 and, depending on their contributions, could move on to the next generations.

In each of $T = 5$ periods, each decision maker receives an endowment of $\frac{1}{T}w$ and decides how much of the endowment to allocate to a “group account” that will exclusively benefit subsequent generations. Thus, when the members of Generation 1 make their decisions, all remaining players are passive players. We denote the contribution amount of player i in round t as $c_{i,t}$. Let $c_i = \sum_{t=1}^T c_{i,t}$ be the total individual contribution of a player i over all T periods, and let $\mathbf{c} = (c_1, \dots, c_9)$ be the vector of the total contributions of all players in generations 1-3. If, at the end of the T periods, the sum of contributions of the three decision makers is equal to or above a threshold value $C = \frac{3}{2}w$, the experiment continues with the next generation whose members become active decision makers. Otherwise, the experiment ends, and the members of the next generation make no decisions and receive no payoff. If the experiment continues with Generation 2 (3), then this generation will face the same decisions as Generation 1 (2). If Generation 3 reaches the threshold, then each member of Generation 4 earns $\frac{1}{2}w$. Otherwise, the game ends with Generation 3, and the members of Generation 4 earn nothing.

We define an indicator function $\mathbb{1}_{G_i}$ which indicates whether or not Generation i has reached the threshold. I.e., we have

$$\mathbb{1}_{G_i}(\mathbf{c}) = \begin{cases} 1, & \text{for } \sum_{j \in G_i} c_j \geq C, \\ 0 & \text{otherwise.} \end{cases}$$

The payoff of player i are thus given by

$$\pi_i(\mathbf{c}) = \begin{cases} w - c_i & \text{for } i \in G_1, \\ (w - c_i)\mathbb{1}_{G_1}(\mathbf{c}) & \text{for } i \in G_2, \\ (w - c_i)\mathbb{1}_{G_1}(\mathbf{c})\mathbb{1}_{G_2}(\mathbf{c}) & \text{for } i \in G_3, \\ \frac{1}{2}w\mathbb{1}_{G_1}(\mathbf{c})\mathbb{1}_{G_2}(\mathbf{c})\mathbb{1}_{G_3}(\mathbf{c}) & \text{for } i \in G_4. \end{cases} \quad (1)$$

In the experimental implementation of the game, we choose parameters such that if each decision maker contributes exactly half of the endowment, the threshold is reached, and all players in the cohort receive equal payoffs ($c_i = \pi_i(c) = \frac{1}{2}w$). Hence, contributing half of the endowment may serve as a focal point for cooperation.

Finally, to see that the IGG is a Game of Altruistic Cooperation, note that for any money-maximizing decision maker (in any generation), there is a dominant strategy of defecting and contributing nothing (Condition a). Moreover, for all decision makers in a given generation, mutual cooperation (i.e., the generation reaches the threshold) results in a lower or equal payoff than mutual defection (i.e., contributing nothing). However, reaching the threshold costs the current generation an amount of $\frac{3}{2}w$ but provides the subsequent generation with an endowment of $3w$ (or $\frac{3}{2}w$ in the case of Generation 4). Thus, mutual cooperation (reaching the threshold) leads to a higher or equal aggregate payoff for all players (including players in future generations) than mutual defection (Condition b*).

The baseline game is summarized in Figure 1. For our experiment, we used $w = 50$ ECU (Experimental currency units). Thus, in each of the five periods, each member of the current generation decides on the allocation of $\frac{1}{5}w = \frac{1}{5}50 = 10$ ECU. The threshold to be reached by a generation is $C = \frac{3}{2}w = 150$ ECU, which can be met if each decision maker contributes half of her endowment (i.e., $c_i = 25$ ECU). If members of all three active generations do this, all participants in the cohort receive equal payoffs ($\pi_i(c) = \frac{1}{2}w = 25$ ECU). If all members of the first generation defect by contributing nothing, they each receive 50 ECU, i.e., a total of 150 ECU, while all remaining players receive nothing. If instead, every generation cooperates by reaching the threshold exactly, total earning across all generations are 300 ECU. In other words, altruistic cooperation maximizes the cohort's aggregate income.

3.1 Treatments

There are four between-subject treatments, crossing the dimensions of punishment and communication in a 2x2 design: BASE (no punishment, no communication - as described in Section 3), PUN (punishment and no communication), PUN&COMMS (punishment and communication), and COMMS (communication and no punishment).

The PUN treatment is identical to BASE, but includes costly peer punishment à la [Fehr and Gächter \(2000, 2002\)](#). After making their contribution

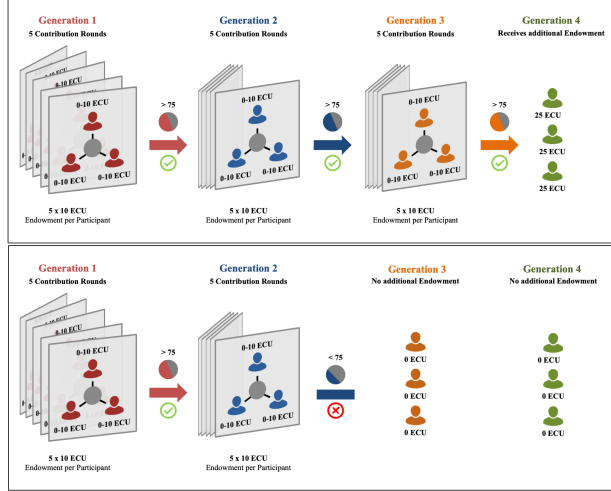


Figure 1: Repeated Intergenerational Goods Game (IGG)

Note: Participants are divided into cohorts of four generations with three members each. In Gen 1, each member is endowed with 10 experimental currency units (ECU) in each of the 5 periods (i.e., a total of 50 ECU per members). In each period, each of the three members simultaneously makes a decision to contribute to a group account. If, at the end of the 5 periods, the members have contributed at least 75 ECU (i.e., half of the total available endowment to the three members), the experiment continues with the next generation, otherwise, the experiment ends. If the contributions of Gen 3 reach the threshold, each member of Gen 4 receives an endowment of 25 ECU, and the game ends without any further decision being made. **Upper panel:** A case with perfect sustained generations (i.e., Gen 4 is reached). **Lower panel:** A case where contributions of Gen 2 have failed to reach the threshold, and the experiment ends. Members of the subsequent Gens 3 and 4 are endowed with 0 ECU.

choices, decision makers observe the individual contributions made by the peers within their generation and can assign 0, 1, or 2 negative points to each of them (as long as their endowment minus their contribution allows for it). Put differently, punishment is contemporaneous i.e., within a generation but not across generations. Each punishment point costs the sender 1 ECU and reduces the payoff of the punished player by 3 ECU. For any generation k and players $i, j \in G_k$, let $p_{ij,t}$ be the number of negative points sent by a player i to a player j in round t . We denote the total effect of punishment on player i (including the fine from receiving punishment and the cost of sending punishment) as

$$p_i = 3 \sum_{t=1}^T \sum_{j \in G_k} p_{ji,t} + \sum_{t=1}^T \sum_{j \in G_k} p_{ij,t}$$

Table 1: Treatment Overview and Number of Participants

| | No Punishment | Punishment |
|------------------|---------------|------------------|
| No Communication | BASE, N=144 | PUN, N=144 |
| Communication | COMMS, N=96 | PUN&COMMS, N=144 |

and $\mathbf{p} = (p_1, \dots, p_9)$. Under punishment, the payoff of player i are given by

$$\pi_i(\mathbf{c}, \mathbf{p}) = \begin{cases} w - c_i - p_i, & \text{for } i \in G_1 \\ (w - c_i - p_i)\mathbb{1}_{G_1}(\mathbf{c}), & \text{for } i \in G_2 \\ (w - c_i - p_i)\mathbb{1}_{G_1}(\mathbf{c})\mathbb{1}_{G_2}(\mathbf{c}), & \text{for } i \in G_3 \\ \frac{1}{2}w\mathbb{1}_{G_1}(\mathbf{c})\mathbb{1}_{G_2}(\mathbf{c})\mathbb{1}_{G_3}(\mathbf{c}), & \text{for } i \in G_4. \end{cases} \quad (2)$$

The COMMS treatment is identical to BASE, except that before the contribution phase, there is a communication phase, in which members of the current generation can chat freely with each other for 3 minutes in periods 1 and 2 and 1.5 minutes in periods 3, 4, and 5 (in an anonymous chat window; i.e., identification is not allowed, (see e.g., [Bochet et al., 2006](#)) and was also not observed in the chat protocols). As before, communication is only possible between members of the same generation and not between members of different generations. The PUN&COMMS treatment combines both punishment and communication. Each period starts with a communication phase (as in COMMS), followed by a contribution phase, and ends with a punishment phase (as in PUN).

3.2 Predictions

Under money-maximizing preferences, contributing nothing is a dominant strategy for all players in the Baseline treatment. Costly peer punishment will not occur if decision makers are money-maximizers, and therefore will not deter decision makers from contributing nothing. Similarly, under money-maximizing preferences, we expect non-binding communication to have no effect on the decision makers' contributions. Since the dominant strategy is to contribute nothing, money-maximizing players would have no reason to propose positive contributions or to follow the suggestions of their generational peers. In summary, under money-maximizing preferences, neither communication nor punishment, nor a combination of the two, will lead to a different outcome than in the baseline: All cohorts will fail to reach the threshold in Generation 1.

However, there is ample evidence that many people are also motivated by “other regarding preferences” that also consider the welfare of others, as we have discussed above in the context of dictator games (see also e.g., [Cooper and Kagel, 2017](#)).⁶ One possible, but not exclusive, candidate to represent such

⁶For instance, [Murphy et al. \(2011\)](#), who measure social value orientation as an empirical operationalization of such preferences, observe that 59% of their participants are classified as *prosocial*, while 35% and 3% are classified as *individualistic* and *competitive* (both consistent with payoff maximizing behavior).

preferences is the inequity-aversion model (Fehr and Schmidt, 1999) under the assumption that players consider their payoff as well as the payoffs of all other players in the cohort.

However, even players with strong inequity-averse preferences will not cooperate if they expect the threshold not to be reached. Thus, to reach the threshold, the three group members must have strong inequity-averse preferences and also believe that their peers share these preferences. This is quite a strong requirement, and therefore, we expect only a small number of generations to reach the threshold in BASE. In COMMS, communication can enhance coordination and strengthen positive beliefs among cooperative players, likely increasing the number of generations reaching the threshold compared to BASE. However, because mutual defection maximizes the current generation’s payoffs, communication in the IGG could also lead to coordination on defection instead of cooperation. Overall, the effect of communication is expected to be at most, small, as communication alone does not allow a player to “force” others to cooperate.

Inequity aversion is especially suitable to explain the effectiveness of punishment in facilitating cooperation (Fehr and Schmidt, 1999). However, unlike in classical social dilemmas, defection in the IGG does not reduce the payoffs of other decision makers (hence, defection is not equivalent to free-riding as typically defined in classical social dilemmas). In PUN&COMMS, those inequity-averse decision makers, who are willing to punish, may reveal this explicitly or implicitly during the communication phase. If this revelation appears credible to their peers, even payoff maximizers could contribute half of their endowment (since the payoff lost to punishment points received is higher than the burden of contributing). Therefore, we expect a positive effect in PUN&COMMS compared to BASE and COMMS. Finally, under punishment without communication (in PUN), the availability of punishment will increase contributions if decision makers believe that one of their peers is willing to punish them. As such beliefs may exist, we expect the level of contributions in PUN to be at least as high as in BASE. Whether contributions are higher in PUN than in the two treatments with communication (COMMS and PUN&COMMS) depends on whether communication affects these beliefs.

3.3 Procedure

The experiment took place between December 2022 and May 2023. Participants were recruited from the experimental laboratories of two major German universities using hroot (Bock et al., 2014). The study was conducted online using oTree (version 5.6) (Chen et al., 2016). A total of 528 participants (55% female) took part in four treatments: BASE (N=144), PUN (N=144), PUN&COMMS (N=144), and COMMS (N=96).

In total, there were 12 independent cohorts of four generations in each of the first three treatments and 8 cohorts in COMMS. All participants received 5 euros as a show-up fee, regardless of their payoff in the game. We conducted 7 sessions per treatment, with one or two cohorts (12 or 24 participants) partici-

pating in each session. Each session consisted of three parts: first, participants received instructions, including quiz questions, after logging into the experiment. The instructions were identical for all participants. Second, participants were randomly assigned to a generation and the experiment began. Finally, participants completed a short post-experiment questionnaire that collected demographic data and information about their motivation and perceived difficulty of the experiment. Participants reported a mean difficulty level of 2.2 (“easy”) with a standard deviation of 0.92 (ranked on a scale from 1 = very easy to 5 = very difficult). Earnings were converted to euros at an ECU/euro exchange rate of 2.5:1 and transferred to participants’ bank accounts within two working days.

4 Results

4.1 Sustained Generations

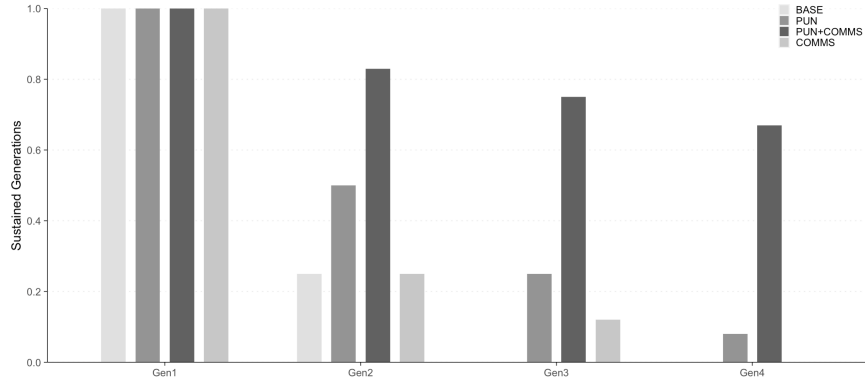


Figure 2: Sustained generations across treatments In BASE, 12 cohorts started in Gen 1. Of these, 3, 0, and 0 cohorts reached Gen 2, Gen 3, and Gen 4, respectively. In PUN, 12 cohorts started in Gen 1. Of these, 6, 3, and 1 cohorts reached Gen 2, Gen 3, and Gen 4. In PUN&COMMS, 12 cohorts started in Gen 1. Of these, 10, 9, and 8 cohorts reached Gen 2, Gen 3, and Gen 4. In COMMS, 8 cohorts started in Gen 1. Of these, 2 and 1 cohorts reached Gen 2 and 3. Then, one participant in Gen 3 dropped out of the experiment, and we had to stop the experiment for this cohort, although the cohort could still have reached Gen 4. Appendix D shows that the inclusion/exclusion of this generation has no significant effect on the results.

Figure 2 shows the percentage of generations that are sustained throughout the game in the different treatments. As the game progresses from one generation to the next, it indicates a very high level of sustained generations in PUN&COMMS, a lower level in PUN, and very low levels of sustained generations in BASE and COMMS. Comparing the effectiveness of the different institutions between treatments, we observe that significantly more generations are sustained in PUN&COMMS than in any of the other treatments (at $p < .05$

in each pairwise comparison; robust rank order test; Appendix, Table B.1).⁷ In BASE, only 25% of the cohorts reached the second generation, and no cohort reached the third (and fourth) generation. The average likelihood of reaching the threshold for a generation in this treatment is 20%. In PUN, 50% reached the second, 25% the third, and 8.3% the fourth generation. These figures closely resemble the results of [Lohse and Waichman \(2020\)](#) - albeit a study with only one period per generation (BASE: 29%, 12%, and 0%; PUN: 56%, 28%, and 12% reached the second, third, and fourth generation, respectively). Thus, punishment increases the number of sustained generations, but only moderately so overall; the average likelihood of reaching the threshold for a generation in this treatment is 48%. However, in contrast to findings on standard public good games (e.g., [Fehr and Gächter, 2000, 2002](#); [Andreoni et al., 2003](#)), the overall success rate of punishment is relatively modest. In fact, in most sequences, punishment does not help sustain the game until generation 4.

In this respect, punishment is significantly more effective when it is combined with communication: in PUN&COMMS, 83%, 75%, and 66% of the cohorts reached generations 2, 3, and 4, respectively. The average likelihood of reaching the threshold for a generation in PUN&COMMS is 87%.

Importantly, in contrast to findings from standard public good games ([Ballet, 2010](#)), communication alone is insufficient to sustain cooperation over multiple generations. In COMMS, only 25% and 12.5% reached generations 2 and 3, respectively. The average likelihood of reaching the threshold for a generation in COMMS is 27%, which is similar to the baseline (20%).

Whether a generation reaches the threshold may depend on its composition, which is randomly determined at the start of the experiment. Therefore, following the literature ([Hauser et al., 2014](#); [Lohse and Waichman, 2020](#)), we complement our results with a bootstrap simulation with 40,000 contribution decisions drawn from the observed contribution decisions in the first period (Appendix, Section E). This simulation provides further evidence that our results are not driven by random generation assignment within a sample of a limited size.

4.2 Individual contributions

Next, we compare individual contributions across treatments, either for the first period (i.e., Period 1 in different cohorts) only or for all periods. The reason for looking at contributions in the first and all periods separately is that in later periods, even those initially inclined to cooperate may stop contributing after observing that the threshold cannot be reached. Contributions in the first period instead reflect the beliefs of participants without having observed the contributions of other decision makers yet.

⁷In addition, the likelihood of sustained generations among all possible generations, PUN&COMMS is significantly larger than in the baseline and any of the other treatments (Chi2 test; $p < 0.001$ for each pairwise comparison between PUN&COMMS and the other treatments; Appendix, Table B.2).

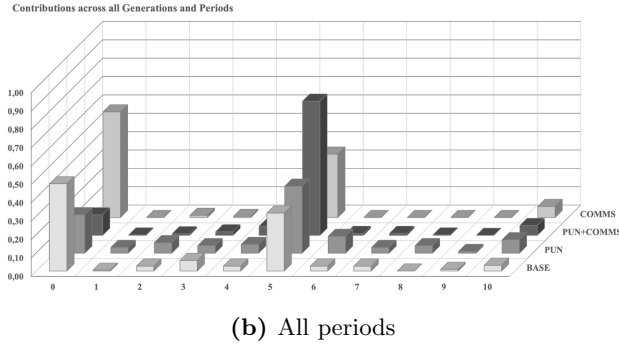
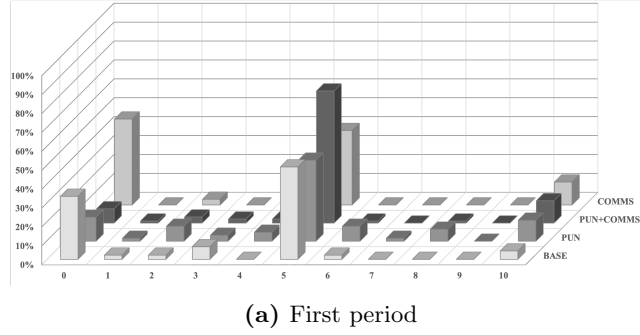


Figure 3: Panel (a): Individual contribution in the first period only (a total of 45, 63, 93, and 33 observations in BASE, PUN, PUN&COMMS, and COMMS). Panel (b): All five periods .

The average contributions are 3.29 and 3.24 in BASE and COMMS. They are markedly higher in PUN and PUN&COMMS with 4.81 and 5.05, respectively. Figure 3a shows that in all treatments except for COMMS, the modal contribution in the first period is 5 ECU. In COMMS, the modal contribution is zero, while a smaller but still considerable share of participants choose the focal contribution level of 5 ECU. This observation shows that communication without punishment may actually undermine cooperation; it allows individuals to highlight their non-cooperative intentions and thereby may discourage potential cooperators from spending on a lost cause.

Table 2 contains coefficient estimates from linear regressions of individual contributions on the set of treatment indicators. Model 1 focuses on the first period of Generation 1, while Model 2 includes the first period of all sustained generations. Models 3 and 4 repeat this analysis using contributions averaged across the five periods as the dependent variable. In all four specifications, communication does not increase contributions relative to the baseline. Both PUN and PUN&COMMS lead to higher contributions compared to BASE and COMMS. In terms of contributions, there is no evidence that PUN&COMMS leads to significantly higher contributions than PUN.

Table 2: Linear Regressions: Individual Contributions

| | Model (1) Period 1: Gen 1 | Model (2) Period 1: All gen | Model (3) All Periods: Gen 1 | Model (4) All Periods: All gen |
|---|------------------------------|--------------------------------|---------------------------------|-----------------------------------|
| COMM | -0.92 (0.89) | -0.05 (0.74) | -0.57 (0.69) | -0.28 (0.56) |
| PUN | 1.06 (0.65) | 1.52*** (0.55) | 1.44*** (0.51) | 1.55*** (0.42) |
| PUN+COMM | 1.72*** (0.58) | 1.76*** (0.48) | 1.97*** (0.43) | 1.97*** (0.37) |
| Constant | 3.42*** (0.48) | 3.29*** (0.41) | 2.73*** (0.38) | 2.65*** (0.33) |
| Observations | 132 | 234 | 132 | 234 |
| R-squared | 0.11 | 0.08 | 0.19 | 0.18 |
| Pairwise treatment comparisons by joint F-tests (p-values) | | | | |
| COMMS vs. PUN | 0.0253** | 0.0277** | 0.0033*** | <0.001*** |
| COMMS vs. PUN&COMMS | 0.0017*** | 0.0064*** | <0.001*** | <0.001*** |
| PUN vs. PUN&COMMS | 0.2256 | 0.5738 | 0.1799 | 0.1674 |

Note: Dependent variable: Individual contributions. Robust standard errors are shown in parentheses. BASE is the default treatment. “Period 1” refers to the first of five periods per generation. “Gen 1” is the first generation, whereas “All gen” indicates individual contributions in all generations that were reached. Finally, ***, **, and * denote significance levels of 1%, 5%, and 10%, respectively (two-sided tests).

The results observed in the first period are generally consistent with those considering decisions in all five periods per generation (see comparison of Figure 3a and 3b; and comparison of models 1 and 3 with 2 and 4 in Table 2). Taking into account all periods, the average individual contributions do not differ between the BASE (average=2.65) and COMMS (2.38) treatments, and between PUN (4.21) and PUN&COMMS (4.63) treatments.

4.3 How do the different institutions drive behavior?

This section provides possible explanations for two puzzles. First, although individual contributions are not significantly higher in PUN&COMMS than in PUN, there is a considerably higher level of sustained generations in the former treatment. Second, why does communication without punishment not lead to a significant improvement over baseline cooperation?

Regarding the first puzzle, in PUN&COMMS, participants use the chat messages to coordinate on the focal contribution level (where symmetric contributions equalize within-generation and across-generations earnings) considerably more often than in PUN. In particular, this focal contribution level occurs at similar frequencies in BASE, COMM, and PUN (32%, 35%, and 37%) and at much higher frequencies in PUN&COMMS (73%). Table 3, uses logit regressions to show that the focal contribution of 5 ECU is significantly more frequent in PUN&COMMS than in any of the other treatments.⁸ This regression also shows that when considering the focal contribution in each of the five periods, then communication enables coordinating on this outcome (significantly higher prob-

⁸Alternative linear probability model regressions yield similar results (see Appendix Table C.2).

ability under COMMS than under either BASE or PUN). Yet, when comparing COMMS to PUN&COMMS the latter has a significantly higher occurrence of focal contributions than COMMS. In the Appendix we repeat the same exercise for defection (zero contributions). Here our results shows that communication is not combined with Punishment, it can actually backfire. Both in the first period or in each of the five periods, COMMS has higher occurrence of zero contributions than the other treatments, while PUN&COMMS has not (see Appendix Table C.5).

Table 3: Logit regressions: Individual contributions of 5 ECU (focal point cooperation)

| | Model (1) Period 1: Gen 1 | Model (2) Period 1: All gen | Model (3) All Periods: Gen 1 | Model (4) All Periods: All gen |
|---|------------------------------|--------------------------------|---------------------------------|-----------------------------------|
| COMM | -1.10* (0.58) | -0.39 (0.47) | 1.73** (0.87) | 1.50** (0.72) |
| PUN | -0.22 (0.47) | -0.24 (0.39) | 0.00 (1.03) | 0.39 (0.74) |
| PUN+COMM | 0.96* (0.50) | 0.89** (0.38) | 3.06*** (0.80) | 3.05*** (0.64) |
| Constant | -0.00 (0.33) | -0.04 (0.30) | -2.83*** (0.73) | -2.64*** (0.60) |
| Observations | 132 | 234 | 132 | 234 |
| Pseudo R-squared | 0.077 | 0.050 | 0.241 | 0.231 |
| Pairwise comparisons of coefficients (Chi2 (1)) (p-values) | | | | |
| COMMS vs. PUN | 0.132 | 0.744 | 0.046** | 0.060* |
| COMMS vs. PUN&COMMS | 0.0007*** | 0.003*** | 0.023** | 0.0007*** |
| PUN vs. PUN&COMMS | 0.019** | 0.0009*** | 0.0001*** | 0.0000*** |

Note: The dependent variable is assigned the value of 1 if a contribution of 5 ECU in the initial period (left two columns) or if contribute 5 ECU in each of the five periods (right two columns), and otherwise zero. Robust standard errors are given in parentheses. BASE is the default treatment. “Period 1” refers to the first of five periods per generation. “Gen 1” is the first generation, whereas “All gen” indicates individual contributions in all generations that were reached. Finally, ***, **, and * denote significance levels of 1%, 5%, and 10%, respectively (two-sided tests).

The importance of this coordinating effect of communication is further highlighted when examining how often generations narrowly fail to reach the threshold between PUN and PUN&COMMS. In PUN, 4 out of 14 generations (29%) contributed at least 90% (68 ECU) of the amount needed to reach the threshold, but narrowly missed it. In PUN&COMMS, only 1 out of 28 generations (4%) narrowly missed it.

To better understand the differences in outcomes between the two communication treatments (with and without punishment), we analyzed the content of the chat transcripts. Four independent annotators evaluated the content of each chat transcript. The annotators were instructed to identify the initial and the final suggestions in each chat protocol and to evaluate whether decision makers suggested to “cooperate”, to “not cooperate”. Unclear cases were marked as “unspecified.”⁹ Initial suggestions may serve as a starting point

⁹In 71% (26%) of the cases there was consensus among annotators’ opinions (three annota-

or anchor for the subsequent discussion, while final suggestions may indicate an ultimate agreement. In PUN&COMMS, initial suggestions to cooperate occurred more often than in COMMS (87% vs. 55%), possibly because non-cooperators were more hesitant about revealing their intentions early on in the chat due to a fear of later punishment. In COMMS, initial suggestions not to contribute occurred in 36% of all chat protocols, compared to only 7% of chat protocols in PUN&COMMS. When comparing initial to final suggestions, in PUN&COMMS, there is virtually no change. In contrast, participants in the COMMS treatment are considerably more likely to suggest making no contributions in their final messages than in the initial message (an increase from 36% to 55%). These final messages are highly predictive of actual behaviour in both treatments: 90% and 97% of generations behaved in line with the final suggestion in COMMS and PUN&COMMS, respectively.

Finally, communication also changes how often punishment occurs when comparing the two punishment treatments. Decision makers rarely make use of the punishment option when communication is possible. In contrast, punishment is applied more often in the absence of communication: 17% of the decisions were punished in PUN, while only 5% were punished in PUN&COMMS (Appendix, Table C.6). This suggests that punishment facilitates cooperation primarily through deterrence (Fudenberg and Pathak, 2010). Interestingly, punished members do not increase their contribution in the following period. The median period-to-period change in an individual’s contribution after being punished is 0 ECU. Thus, even when applied, punishment has no disciplinary effect in the following period.

The sparse use of punishment in PUN&COMMS is consistent with the corresponding chat protocols. The annotators observed only a few instances (6%) in which participants explicitly threatened others with punishment. More common were instances where participants reassured their generational peers that they would not punish each other if they adhered to the agreed-upon suggestion (19%). Thus, in PUN&COMMS, the implicit threat of punishment, rather than an explicit reference to it in the chat or subsequent enforcement, is sufficient to achieve high levels of sustained generations.

Finally, when comparing average payoffs within the cohorts between treatments, we observe that the average payoffs are substantially higher in PUN&COMMS than in any other treatment (robust rank order test; $p < 0.05$ for each pairwise comparison). This outcome is driven by (i) a higher proportion of sustained generations in PUN&COMMS, (ii) few resources spent on punishment, and (iii) improved coordination. The latter enables decision makers to either reach the threshold without wasting resources, or to make an early decision to avoid wasting resource when there is insufficient chance for altruistic cooperation to succeed. That is, improved coordination avoids narrowly missing the threshold and wasting resources in vain.

tors had the same opinion). In 3% of the instances (labeled as "unspecified") two annotators had the same opinion.

5 Discussion

Games of Altruistic Cooperation model situations in which a group of decision makers must sacrifice some payoff for the exclusive benefit of others. The dilemma arises because mutual cooperation maximizes the aggregate income of all players, while mutual defection is best not only for the individual but also for the group of decision makers. Using the IGG as an example of this class of games, we find that common institutions that are effective in classical social dilemmas — namely, peer punishment and communication — are not sufficient to ensure a high level of altruistic cooperation. Only the combination of these decentralized institutions achieves high levels of altruistic cooperation.

These findings highlight the importance of recognizing that the incentive structure inherent in Games of Altruistic Cooperation is fundamentally different from that in classical social dilemma. Examples of such situations abound in an interconnected economy, where the benefits of public goods provision are rarely limited in time or space to those who decide to fund them. Both the theoretical considerations and the corresponding experimental evidence we provide here suggest that findings from classical social dilemmas may not directly transfer to situations where the beneficiaries are strangers outside of the group of decision makers or future generations as in the case of climate change. Such transfers may overlook key behavioral factors that are unique to altruistic cooperation. This is in line with previous studies showing that contributions to climate change mitigation are not strongly correlated with behavior in standard public good games, suggesting that different underlying motivations are at play (Goeschl et al., 2020). For a general discussion on the transferability of evidence from one experimental context to another see List (2020).

Although stylized in nature, studying Games of Altruistic Cooperation such as the IGG provides valuable insights into climate change and biodiversity loss. While centralized coordination and enforcement are effective in providing such public goods, these institutions are often politically challenging and infeasible at the international level. We provide the first evidence that decentralized institutions, which are more feasible at the international level, can largely solve the intergenerational good dilemma. One interpretation is that international climate diplomacy, which has largely been constructed around communication (e.g., COP meetings), is unlikely to solve this dilemma. Our findings point to the need to address these challenges through appropriate institutional design.

This paper assesses how effectively decentralized institutions promote cooperation in the IGG, treated here as a benchmark case of altruistic cooperation. Future research should examine the portability of these institutions to other environments characterized by altruistic cooperation, thereby testing whether their efficacy depends on specific IGG features like its generational sequence or contribution threshold. Relatedly, alternative experimental designs for studying intergenerational public goods provision could probe the significance of factors like generational structure (Freitas-Groff et al., 2025) or asymmetric impacts on different generations (Hauser, 2025). All these scenarios encapsulate the core problem of altruistic cooperation: a strategic incentive for the present genera-

tion to jointly defect at the expense of future generations.

References

- Andreoni, J., Harbaugh, W., and Vesterlund, L. (2003). The carrot or the stick: Rewards, punishments, and cooperation. *American Economic Review*, 93(3):893–902.
- Axelrod, R. (1981). The evolution of cooperation. *Science*, 212(4489):1390–1396.
- Balliet, D. (2010). Communication and cooperation in social dilemmas: A meta-analytic review. *Journal of Conflict Resolution*, 54(1):39–57.
- Balmford, B., Marino, M., and Hauser, O. P. (2024). Voting sustains intergenerational cooperation, even when the tipping point threshold is ambiguous. *Environmental and Resource Economics*, 87(1):167–190.
- Bardsley, N. (2008). Dictator game giving: altruism or artefact? *Experimental Economics*, 11:122–133.
- Beckenkamp, M. (2006). A game-theoretic taxonomy of social dilemmas. *Central European Journal of Operations Research*, 14(3):337–353.
- Becker, B. (2015). Public r&d policies and private R&D investment: A survey of the empirical evidence. *Journal of Economic Surveys*, 29(5):917–942.
- Bergstrom, T., Blume, L., and Varian, H. (1986). On the private provision of public goods. *Journal of Public Economics*, 29(1):25–49.
- Bicchieri, C. (2002). Covenants without swords: Group identity, norms, and communication in social dilemmas. *Rationality and Society*, 14(2):192–228.
- Bochet, O., Page, T., and Putterman, L. (2006). Communication and punishment in voluntary contribution experiments. *Journal of Economic Behavior & Organization*, 60(1):11–26.
- Bochet, O. and Putterman, L. (2009). Not just babble: Opening the black box of communication in a voluntary contribution experiment. *European Economic Review*, 53(3):309–326.
- Bock, O., Baetge, I., and Nicklisch, A. (2014). hroot: Hamburg registration and organization online tool. *European Economic Review*, 71:117–120.
- Bonatti, A. and Hörner, J. (2011). Collaborating. *American Economic Review*, 101(2):632–663.
- Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., et al. (2018). Evaluating the replicability of social science experiments in nature and science between 2010 and 2015. *Nature Human Behaviour*, 2(9):637–644.

- Cason, T. N. and Gangadharan, L. (2016). Swords without covenants do not lead to self-governance. *Journal of Theoretical Politics*, 28(1):44–73.
- Cason, T. N. and Mui, V.-L. (1997). A laboratory study of group polarisation in the team dictator game. *The Economic Journal*, 107(444):1465–1483.
- Chan, D. C. (2016). Teamwork and moral hazard: evidence from the emergency department. *Journal of Political Economy*, 124(3):734–770.
- Chen, D. L., Schonger, M., and Wickens, C. (2016). oTree—an open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, 9:88–97.
- Cooper, D. J. and Kagel, J. H. (2017). *Other-Regarding Preferences: A Selective Survey of Experimental Results*, pages 217–289. Princeton University Press, Princeton.
- Dawes, R. M. (1980). Social dilemmas. *Annual Review of Psychology*, 31(1):169–193.
- Dawes, R. M. and Thaler, R. H. (1988). Anomalies: Cooperation. *Journal of Economic Perspectives*, 2(3):187–197.
- Delaney, J. and Jacobson, S. (2014). Those outsiders: How downstream externalities affect public good provision. *Journal of Environmental Economics and Management*, 67(3):340–352.
- Engel, C. (2011). Dictator games: A meta study. *Experimental Economics*, 14:583–610.
- Fehr, E. and Gächter, S. (2000). Cooperation and punishment in public goods experiments. *American Economic Review*, 90(4):980–994.
- Fehr, E. and Gächter, S. (2002). Altruistic punishment in humans. *Nature*, 415(6868):137–140.
- Fehr, E. and Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics*, 114(3):817–868.
- Forsythe, R., Horowitz, J. L., Savin, N. E., and Sefton, M. (1994). Fairness in simple bargaining experiments. *Games and Economic behavior*, 6(3):347–369.
- Freitas-Groff, Z., Grodeck, B., Hauser, O., and Lohse, J. (2025). Cooperating across generations: Reciprocal cooperation or intergenerational exchange? *Available at SSRN*.
- Fudenberg, D. and Pathak, P. A. (2010). Unobserved punishment supports cooperation. *Journal of Public Economics*, 94(1-2):78–86.
- Gardner, R., Ostrom, E., and Walker, J. M. (1990). The nature of common-pool resource problems. *Rationality and Society*, 2(3):335–358.

- Goeschl, T., Kettner, S. E., Lohse, J., and Schwieren, C. (2020). How much can we learn about voluntary climate action from behavior in public goods games? *Ecological Economics*, 171:106591.
- Hardin, G. (1968). The tragedy of the commons: the population problem has no technical solution; it requires a fundamental extension in morality. *Science*, 162(3859):1243–1248.
- Hauser, O. P. (2025). Climate change, intergenerational fairness, and the promises and pitfalls of artificial intelligence. *Environmental and Resource Economics*, pages 1–18.
- Hauser, O. P., Rand, D. G., Peysakhovich, A., and Nowak, M. A. (2014). Cooperating with the future. *Nature*, 511(7508):220–223.
- Hirshleifer, J. (1978). Competition, cooperation, and conflict in economics and biology. *American Economic Review*, 68(2):238–243.
- Isaac, R. M. and Walker, J. M. (1988a). Communication and free-riding behavior: The voluntary contribution mechanism. *Economic Inquiry*, 26(4):585–608.
- Isaac, R. M. and Walker, J. M. (1988b). Group size effects in public goods provision: The voluntary contributions mechanism. *Quarterly Journal of Economics*, 103(1):179–199.
- Janssen, M. A., Holahan, R., Lee, A., and Ostrom, E. (2010). Lab experiments for the study of social-ecological systems. *Science*, 328(5978):613–617.
- Kamei, K., Putterman, L., and Tyran, J.-R. (2015). State or nature? endogenous formal versus informal sanctions in the voluntary provision of public goods. *Experimental Economics*, 18:38–65.
- Karlan, D. and List, J. A. (2007). Does price matter in charitable giving? evidence from a large-scale natural field experiment. *American Economic Review*, 97(5):1774–1793.
- Kosfeld, M. and Rustagi, D. (2015). Leader punishment and cooperation in groups: Experimental field evidence from commons management in ethiopia. *American Economic Review*, 105(2):747–783.
- Kreps, D. M., Milgrom, P., Roberts, J., and Wilson, R. (1982). Rational cooperation in the finitely repeated prisoners’ dilemma. *Journal of Economic Theory*, 27(2):245–252.
- Ledyard, J. O. (1994). Public goods: A survey of experimental research. In *The Handbook of Experimental Economics*, pages 111–194. Princeton University Press.
- List, J. A. (2007). On the interpretation of giving in dictator games. *Journal of Political Economy*, 115(3):482–493.

- List, J. A. (2020). Non est disputandum de generalizability? a glimpse into the external validity trial. Technical report, National Bureau of Economic Research.
- Lohse, J. and Waichman, I. (2020). The effects of contemporaneous peer punishment on cooperation with the future. *Nature Communications*, 11(1):1815.
- Luhan, W. J., Kocher, M. G., and Sutter, M. (2009). Group polarization in the team dictator game reconsidered. *Experimental Economics*, 12:26–41.
- Markussen, T., Putterman, L., and Tyran, J.-R. (2014). Self-organization for collective action: An experimental study of voting on sanction regimes. *The Review of Economic Studies*, pages 301–324.
- Murphy, R. O., Ackermann, K. A., and Handgraaf, M. J. J. (2011). Measuring social value orientation. *Judgment and Decision Making*, 6(8):771–781.
- Olson, M. (1965). *The Logic of Collective Action: Public Goods and the Theory of Groups, with a new preface and appendix*. Harvard University Press, Cambridge, MA.
- Ostrom, E., Gardner, R., and Walker, J. (1994). *Rules, Games, and Common-Pool Resources*. University of Michigan press.
- Ostrom, E., Walker, J., and Gardner, R. (1992). Covenants with and without a sword: Self-governance is possible. *American Political Science Review*, 86(2):404–417.
- Palfrey, T., Rosenthal, H., and Roy, N. (2017). How cheap talk enhances efficiency in threshold public goods games. *Games and Economic Behavior*, 101:234–259.
- Panchanathan, K., Frankenhuis, W. E., and Silk, J. B. (2013). The bystander effect in an n-person dictator game. *Organizational Behavior and Human Decision Processes*, 120(2):285–297.
- Rand, D. G. and Nowak, M. A. (2013). Human cooperation. *Trends in Cognitive Sciences*, 17(8):413–425.
- Roelfsema, M., van Soest, H. L., Harmsen, M., van Vuuren, D. P., Bertram, C., den Elzen, M., Höhne, N., et al. (2020). Taking stock of national climate policies to evaluate implementation of the Paris agreement. *Nature Communications*, 11(1):2096.
- Schelling, T. C. (1973). Hockey helmets, concealed weapons, and daylight saving: A study of binary choices with externalities. *Journal of Conflict Resolution*, 17(3):381–428.
- Smith, E. A. (2010). Communication and collective action: language and the evolution of human cooperation. *Evolution and Human Behavior*, 31(4):231–245.

- Tavoni, A., Dannenberg, A., Kallis, G., and Löschel, A. (2011). Inequality, communication, and the avoidance of disastrous climate change in a public goods game. *Proceedings of the National Academy of Sciences*, 108(29):11825–11829.
- Vesterlund, L. (2016). Using experimental methods to understand why and how we give to charity. *Handbook of experimental economics*, 2(91):151.
- Villeval, M. C. (2020). Public goods, norms and cooperation. In *Handbook of Experimental Game Theory*, pages 184–212. Edward Elgar Publishing.
- Weiss, E. B. (1990). Our rights and obligations to future generations for the environment. *American Journal of International Law*, 84(1):198–207.
- Zelmer, J. (2003). Linear public goods experiments: A meta-analysis. *Experimental Economics*, 6:299–310.

A Experimental design and procedure: Extension

A.1 Experimental design

We begin by describing our terminology. Treatment is the experimental condition (BASE, PUN, PUN&COMMS, or COMMS), and Session refers to a specific date and time. Each session consists of either one or two cohorts. One cohort consists of 12 participants divided into 4 generations of 3 members. In each generation, members make decisions in 5 consecutive periods.

In the original version of the intergenerational good game (IGG), decision makers in each generation make a single, one-shot decision, with an indefinite number of generations (Hauser et al., 2014). Later adaptations retain the feature of one-shot decisions, but limit the game to a finite number of generations (Lohse and Waichman, 2020).¹ The primary goal of our study is to compare the effectiveness of decentralized punishment- and communication-based institutions in sustaining future generations. To this end, we introduce a further modification to the original IGG design, by incorporating intragenerational repetition (i.e., five decision periods per generation). After each period, there is a feedback screen showing the contributions made by each decision maker (in the current generation). This modification allows us to gauge the full effectiveness of each institution, as decision-makers can observe how others in their generation use and react to institutions such as peer-punishment.²

The design of the baseline game is described in Figure A.1. Each participant receives a €5.00 show-up fee and can earn additional money depending on whether his/her generation is reached and his/her decisions (and in PUN and PUN&COMMS, also on the punishment decisions of their generational peers). In BASE, in each period, each member of Gen 1 is endowed with 10 ECU (Experimental Currency Units; the exchange rate in the experiment is 2.5:1, i.e., 2.5 ECU are exchanged for €1.00) per period. I.e., each member receives 50 ECU in total (10 ECU in each period for five periods). In every period, each member decides how many ECU to allocate to a group account (while the rest of the endowment remains (automatically) in their private account). If the sum of all individual contributions to the group account at the end of the five periods is less than 75 ECU, the experiment ends instantly, and the payoff for all participants of later generations is zero. If the sum of all individual contributions to the group account amounts to 75 ECU or more (i.e., at least 50% of the total generation endowment), the game continues to Gen 2. Independent from the decisions of later generations, members of Gen 1 earn the respective amounts

¹The similarity in baseline cooperation rates between the two approaches (Lohse and Waichman, 2020) suggests that the number of generations is not a crucial design element.

²Allowing multiple decision periods per generation reduces the financial risk of cooperation and may increase altruistic cooperation (Schelling, 1960). Additionally, repetition can enhance the impact of punishment by allowing for strategic motives, and has been deliberately used in modelling climate change negotiations within intragenerational settings (Milinski et al., 2008; Tavoni et al., 2011).

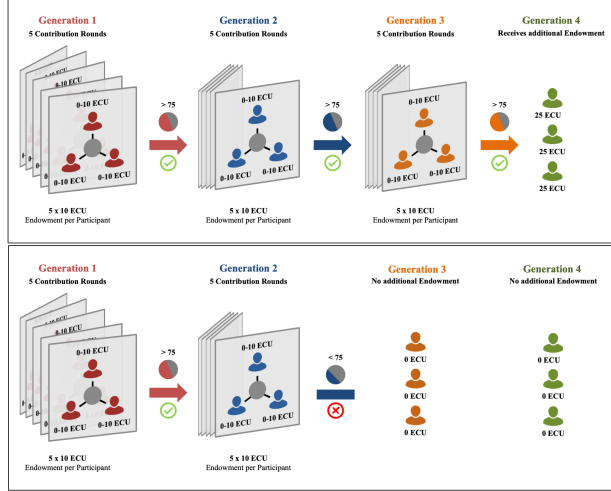


Figure A.1: Repeated Intergenerational Goods Game (IGG)

Note: Participants are divided into cohorts of four generations with three members each. In Gen 1, each member is endowed with 10 experimental currency units (ECU) in each of the 5 periods (i.e., a total of 50 ECU per members). In each period, each of the three members simultaneously makes a decision to contribute to a group account. If, at the end of the 5 periods, the members have contributed at least 75 ECU (i.e., half of the total available endowment to the three members), the experiment continues with the next generation, otherwise, the experiment ends. If the contributions of Gen 3 reach the threshold, each member of Gen 4 receives an endowment of 25 ECU, and the game ends without any further decision being made. **Upper panel:** A case with perfect sustained generations (i.e., Gen 4 is reached). **Lower panel:** A case where contributions of Gen 2 have failed to reach the threshold, and the experiment ends. Members of the subsequent Gens 3 and 4 are endowed with 0 ECU.

in their private accounts (i.e., the total endowment of 50 ECU minus the sum of their individual contributions to the group account in the 5 periods). The same procedure applies to members of Gens 2 and 3. The experiment ends with the decisions of Gen 3. If they contribute 75 ECU or more, then each member of Gen 4 earns 25 ECU. Otherwise, members of Gen 4 earn nothing. For Gens 1 and 2, reaching the threshold costs 75 ECU to the current generation but provides an endowment of 150 ECU to the subsequent generation(s). Hence, there is a gain in total earnings (across multiple generations) from reaching the threshold. We choose the parameters such that there is a symmetric focal point for cooperation (each member contributes a total of 25 ECU in the five periods).

A.2 Procedure and treatments

We conducted 28 sessions between December 7, 2022, and May 8, 2023. Of these, 12 sessions consisted of one, and 16 sessions consisted of two cohorts. We treat each cohort as an independent observation. After running 22 sessions with randomized order of the three treatments BASE, PUN, and PUN&COMMS,

we found that the likelihood of sustained generations is considerably higher in PUN&COMMS than in BASE and PUN. At this point, we could not disentangle whether this result is because of the combined effects of punishment and communication or whether it is due to communication alone. Therefore, we added COMMS as a separate treatment, and Sessions 20 and 23-28 included only the COMMS treatment. We conducted the experiment online using oTree version 5.6 (Chen et al., 2016). The 528 participants were recruited using hroot (Bock et al., 2014) from two large German universities’ experimental economics participant pools. Participants belonged to various faculties (thereof 116 economic students). 340 participants were from one university and 188 from the other university. In total, 17 and 7 sessions included students from either the first or the second university, and 4 sessions included students from both universities.

| Treatment | Sessions | Cohorts | Participants | Avg. Payout (€) | Avg. Duration (Mins) |
|-----------|----------|---------|--------------|-----------------|----------------------|
| BASE | 7 | 12 | 144 | 9.59 | 27.86 |
| PUN | 7 | 12 | 144 | 9.44 | 34.86 |
| PUN&COMMS | 7 | 12 | 144 | 13.25 | 59.14 |
| COMMS | 7 | 8 | 96 | 10.24 | 35.00 |
| Total | 28 | 44 | 528 | 10.67 | 39.21 |

Note: “Sessions” = experimental date and time (includes 1 or 2 cohorts), “Cohort” = independent sequence of 12 participants divided into four generations. Average Payout includes a show-up fee of €5.00. “Avg. Duration (Mins)” the time from logging in to login out of the experiment (this includes the instructions)

Table A.1: Treatments

Each session consisted of three phases: First, participants received instructions (including quiz questions) upon logging in to the experiment. The instructions were identical for all participants. Second, participants were randomly assigned to a specific generation, and the experiment started. After making their contribution decision in each period, each member saw a screen with his/her contributions, the contributions of the two other members in their generation, the sum of the periods’ contributions (all these for the current and past period), and the total contribution up to this point.

Each period of the COMMS and PUN&COMMS treatments includes a ‘communication phase’ where members of the current generation could communicate via a chat window before making contribution decisions. They had 3 minutes for communication in the initial two periods and 1.5 minutes in the final three. Participants were not allowed to identify themselves in the chat (see e.g., Bochet et al., 2006). After making the contribution decisions and learning about the individual contribution decisions, each period of the PUN and PUN&COMMS treatments includes a ‘punishment phase’ where members of the current generation could assign 0, 1, or at most 2 negative points to each of the two other members in their generation after learning about the individual contribution decisions. Each negative point costs 1 ECU to the punisher and reduces the punishee’s payoff by 3 ECU. After the punishment decision, members of this current generation received feedback on the total of punishment points received.

We had 28 dropouts (i.e., participants not finishing the experiment). In

compliance with ethics procedures, participants were informed that they could leave the experiment at any time without giving any reason. As we conducted the experiment online, some participants might have experienced technical difficulties and dropped out. Others may have been frustrated waiting and left. In all but one case, the drop-outs were from generations that were not reached. Thus, these dropouts do not affect the performance in the experiment. However, in one cohort of COMMS in Gen 3, one of the active participants dropped out during the experiment. This particular cohort reached Gen 3, but because of the dropout, it could not continue (we know that the cohort reached Gen 3 and could have reached Gen 4). As shown in Section D, this does not affect the statistical analysis (as all other 7 cohorts in COMMS failed in Gen 1).

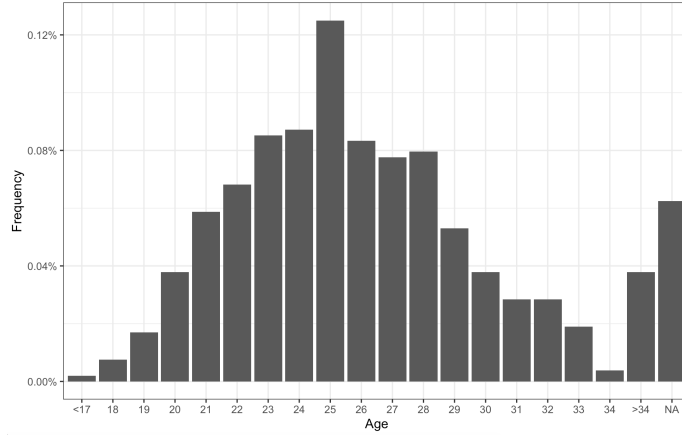


Figure A.2: Age distribution of participants

After all participants in the session had made their decisions, they were asked to complete a short post-experimental questionnaire that provided demographic data and information about their motivation and perceived difficulty of the experiment. Out of 528 participants in the experiment, 33 did not fill in the questionnaire (28 drop-outs and an additional 5 who had technical problems with the questionnaire and contacted the experimenter). Regarding gender, of the 495 participants who filled in the questionnaire, 55.35%, 43.23%, and 1.41% self-identified their gender as female, male, and diverse, respectively. In particular, of the 495 participants, one indicated an age of “less than 17,” and 20 indicated an age of “larger than 34.” The remaining 474 participants self-reported an average age of 25.44 years (std. dev = 3.44). For more details, see the distribution in Figure A.2. As to the subject of study, out of the 495 observations, 23.43% reported studying economics/business. Finally, participants indicated, on average, a difficulty level of 2.2 (“Easy”), with a standard deviation of 0.92 (ranked on a scale from 1=Very Easy to 5=Very Difficult). The average duration of one session was 39 minutes. The average payout was €10.67 per participant (minimum was €5.00 and maximum €25.00). We transferred

the earnings to the participants' bank accounts within 2 days.

B Pairwise tests

In this section, we provide details on the pairwise tests reported in the article.

Table B.1: Occurrence of sustained generations in the different treatments

| | Observations | U-statistic |
|----------------------|--------------|-------------|
| BASE vs. COMMS: | 12 vs. 8 | -0.106 |
| BASE vs. PUN: | 12 vs. 12 | -1.299 |
| BASE vs. PUN&COMMS: | 12 vs. 12 | -3.774*** |
| COMMS vs. PUN: | 8 vs. 12 | -0.974 |
| COMMS vs. PUN&COMMS: | 8 vs. 12 | -3.321** |
| PUN vs. PUN&COMMS: | 12 vs. 12 | -2.743** |

Note: Two-sample Fligner-Policello robust rank order tests. The dependent variable, *Occurrence of sustained generations*, ranges between 1-4, indicating the number of sustained generations reached per cohort. *, **, and *** denote 10%, 5%, and 1% significance levels (two-sided small-sample using [Feltovich, 2005](#)).

Table B.2: Success rate in sustaining generations across all possible generations

| | Observations | $\chi^2(1)$ -statistic | P-value (two-sided) |
|----------------------|--------------|------------------------|---------------------|
| BASE vs. COMMS: | 36 vs. 24 | 0.2778 | 0.598 |
| BASE vs. PUN: | 36 vs. 36 | 4.5997** | 0.032 |
| BASE vs. PUN&COMMS: | 36 vs. 36 | 32.9143*** | 0.000 |
| COMMS vs. PUN: | 24 vs. 36 | 1.9804 | 0.159 |
| COMMS vs. PUN&COMMS: | 24 vs. 36 | 22.5000*** | 0.000 |
| PUN vs. PUN&COMMS: | 36 vs. 36 | 16.06800*** | 0.000 |

Note: Pairwise Chi-squared tests. In creating the dependent variable, *Success rate in sustaining generations*, We assume that any cohort has possible 3 generations. For example, if a cohort reached Gen 4, the success is (1,1,1), whereas if a cohort failed in Gen 1, the data are recorded as (0,0,0). Finally, *, **, and *** denote 10%, 5%, and 1% significance levels.

Table B.3: Total payoffs per cohort

| | Observations | U-statistic |
|----------------------|--------------|-------------|
| BASE vs. COMMS: | 12 vs. 8 | -0.909 |
| BASE vs. PUN: | 12 vs. 12 | 0.534 |
| BASE vs. PUN&COMMS: | 12 vs. 12 | -3.714*** |
| COMMS vs. PUN: | 8 vs. 12 | 0.991 |
| COMMS vs. PUN&COMMS: | 8 vs. 12 | -2.796** |
| PUN vs. PUN&COMMS: | 12 vs. 12 | -4.466*** |

Note: Two-sample Fligner-Policello robust rank order tests. The dependent variable, *Total payoffs per cohort*, is the total payoffs per cohort. *, **, and *** denote 10%, 5%, and 1% significance levels (two-sided small-sample using [Feltovich, 2005](#)).

C Regressions and their tests

In this section, we provide details on the regressions and coefficient tests reported in the article.

Table C.1: Linear regressions: Individual contributions

| | Model 1 Period 1: Gen 1 | Model 2 Period 1: All gen | Model 3 All Periods: Gen 1 | Model 4 All Periods: All gen |
|--------------------------------------|------------------------------------|------------------------------------|------------------------------------|------------------------------------|
| COMMS | -0.92 (0.89) | -0.05 (0.74) | -0.57 (0.69) | -0.28 (0.56) |
| PUN | 1.06 (0.65) | 1.52*** (0.55) | 1.44*** (0.51) | 1.55*** (0.42) |
| PUN&COMMS | 1.72*** (0.58) | 1.76*** (0.48) | 1.97*** (0.43) | 1.97*** (0.37) |
| Constant | 3.42*** (0.48) | 3.29*** (0.41) | 2.73*** (0.38) | 2.65*** (0.33) |
| Observations | 132 | 234 | 132 | 234 |
| R-squared | 0.11 | 0.08 | 0.19 | 0.18 |
| Pairwise comparisons of coefficients | | | | |
| BASE vs. COMMS | t-statistics = -1.03, p = 0.307 | t-statistics = -0.06, p = 0.950 | t-statistics = -0.82, p = 0.415 | t-statistics = -0.50, p = 0.619 |
| BASE vs. PUN | t-statistics = 1.62, p = 0.108 | t-statistics = 2.78, p = 0.006 | t-statistics = 2.83, p = 0.005 | t-statistics = 3.69, p = 0.000 |
| BASE vs. PUN&COMMS | t-statistics = 2.95, p = 0.004 | t-statistics = 3.67, p = 0.000 | t-statistics = 4.56, p = 0.000 | t-statistics = 5.38, p = 0.000 |
| COMMS vs. PUN | F(1, 128) = 5.13, p = 0.0253 | F(1, 230) = 4.91, p = 0.0277 | F(1, 128) = 8.98, p = 0.0033 | F(1, 230) = 12.53, p = 0.0005 |
| COMMS vs. PUN&COMMS | F(1, 128) = 10.31, p = 0.0017 | F(1, 230) = 7.57, p = 0.0064 | F(1, 128) = 17.08, p = 0.0001 | F(1, 230) = 22.59, p = 0.0000 |
| PUN vs. PUN&COMMS | F(1, 128) = 1.48, p = 0.2256 | F(1, 230) = 0.32, p = 0.5738 | F(1, 128) = 1.82, p = 0.1799 | F(1, 230) = 1.92, p = 0.1674 |

Note: The dependent variable is the contribution in the initial period (left two columns) or averaged over the five periods (right two columns). Robust stand errors are given in parentheses. BASE is the default treatment. *, **, and *** denote 10%, 5%, and 1% significance levels (p-values are two-sided).

Table C.2: Linear probability model regressions: Individual contributions of 5 ECU (focal point cooperation)

| | Model 1 Period 1: Gen 1 | Model 2 Period 1: All gen | Model 3 All Periods: Gen 1 | Model 4 All Periods: All gen |
|--------------------------------------|------------------------------------|------------------------------------|------------------------------------|-----------------------------------|
| COMMS | -0.25** (0.13) | -0.09 (0.12) | -0.05 (0.10) | 0.03 (0.09) |
| PUN | -0.06 (0.12) | -0.06 (0.10) | -0.02 (0.07) | 0.05 (0.06) |
| PUN&COMMS | 0.22* (0.12) | 0.21** (0.09) | 0.38*** (0.08) | 0.42*** (0.06) |
| Constant | 0.50*** (0.09) | 0.49*** (0.08) | 0.33*** (0.05) | 0.32*** (0.05) |
| Observations | 132 | 234 | 132 | 234 |
| R-squared | 0.10 | 0.07 | 0.21 | 0.22 |
| Pairwise comparisons of coefficients | | | | |
| BASE vs. COMMS | t-statistics = -1.99, p = 0.049 | t-statistics = -0.82, p = 0.415 | t-statistics = -0.48, p = 0.630 | t-statistics = 0.35, p = 0.729 |
| BASE vs. PUN | t-statistics = -0.46, p = 0.646 | t-statistics = -0.61, p = 0.543 | t-statistics = -0.22, p = 0.823 | t-statistics = 0.77, p = 0.440 |
| BASE vs. PUN&COMMS | t-statistics = 1.93, p = 0.056 | t-statistics = 2.33, p = 0.021 | t-statistics = 4.62, p = 0.000 | t-statistics = 6.87, p = 0.000 |
| COMMS vs. PUN | F(1, 128) = 2.40, p = 0.1239 | F(1, 230) = 0.10, p = 0.7492 | F(1, 128) = 0.10, p = 0.7486 | F(1, 230) = 0.05, p = 0.8171 |
| COMMS vs. PUN&COMMS | F(1, 128) = 15.48, p = 0.0001 | F(1, 230) = 9.30, p = 0.0026 | F(1, 128) = 15.31, p = 0.0001 | F(1, 230) = 22.06, p = 0.0000 |
| PUN vs. PUN&COMMS | F(1, 128) = 5.87, p = 0.0168 | F(1, 230) = 11.55, p = 0.0008 | F(1, 128) = 23.17, p = 0.0000 | F(1, 230) = 39.25, p = 0.0000 |

Note: The dependent variable is the contribution of 5 ECU in the initial period (left two columns) or averaged over the five periods (right two columns). Robust standard errors are given in parentheses. BASE is the default treatment. *, **, and *** denote 10%, 5%, and 1% significance levels (p-values are two-sided).

Table C.3: Logit model regressions: Individual contributions of 5 ECU (focal point cooperation)

| | Model 1 Period 1: Gen 1 | Model 2 Period 1: All gen | Model 3 All Periods: Gen 1 | Model 4 All Periods: All gen |
|---|----------------------------|------------------------------|-------------------------------|---------------------------------|
| COMM | -1.099* (0.580) | -0.386 (0.466) | -1.157** (0.552) | -0.490 (0.478) |
| PUN | -0.223 (0.475) | -0.243 (0.393) | 0.135 (0.521) | 0.045 (0.424) |
| PUN&COMM | 0.956* (0.501) | 0.887** (0.375) | 1.577** (0.706) | 1.439*** (0.477) |
| Constant | -0.000 (0.335) | -0.044 (0.299) | 0.821** (0.363) | 0.795** (0.323) |
| Observations | 132 | 234 | 132 | 234 |
| Pseudo R-squared | 0.077 | 0.050 | 0.116 | 0.080 |
| Pairwise comparisons of coefficients (Chi2 (1)) | | | | |
| BASE vs. COMMS | 3.59, p = 0.058 | 0.69, p = 0.0407 | 4.40, p = 0.036 | 1.05, p = 0.306 |
| BASE vs. PUN | 0.22, p = 0.638 | 0.38, p = 0.536 | 0.07, p = 0.796 | 0.01, p = 0.916 |
| BASE vs. PUN&COMMS | 3.63, p = 0.0567 | 5.59, p = 0.018 | 4.99, p = 0.025 | 9.09, p = 0.0026 |
| COMMS vs. PUN | 2.27, p = 0.132 | 0.11, p = 0.744 | 5.35, p = 0.029 | 1.43, p = 0.233 |
| COMMS vs. PUN&COMMS | 11.61, p = 0.0007 | 9.06, p = 0.0026 | 13.87, p = 0.0002 | 14.98, p = 0.0001 |
| PUN vs. PUN&COMMS | 5.49, p = 0.019 | 10.97, p = 0.0009 | 4.11, p = 0.0426 | 9.75, p = 0.0018 |

Note: The dependent variable is the contribution of 5 ECU in the initial period (left two columns) or averaged over the five periods (right two columns). Robust standard errors are given in parentheses. BASE is the default treatment. *, **, and *** denote 10%, 5%, and 1% significance levels (p-values are two-sided).

Table C.4: Linear probability model regressions: Individual contributions of 0 ECU

| | Model 1 | Model 2 | Model 3 | Model 4 |
|--------------------------------------|------------------------------------|------------------------------------|------------------------------------|------------------------------------|
| | Period 1: Gen 1 | Period 1: All gen | All Periods: Gen 1 | All Periods: All gen |
| BASE | -0.29** (0.13) | -0.12 (0.11) | -0.16 (0.11) | -0.10 (0.10) |
| PUN | -0.51*** (0.12) | -0.33*** (0.10) | -0.47*** (0.11) | -0.37*** (0.09) |
| PUN&COMMS | -0.60*** (0.11) | -0.38*** (0.09) | -0.53*** (0.10) | -0.46*** (0.08) |
| Constant | 0.62*** (0.10) | 0.45*** (0.09) | 0.64*** (0.10) | 0.58*** (0.08) |
| Observations | 132 | 234 | 132 | 234 |
| R-squared | 0.25 | 0.13 | 0.29 | 0.23 |
| Pairwise comparisons of coefficients | | | | |
| COMMS vs. BASE | t-statistics = -2.23, p = 0.028 | t-statistics = -1.06, p = 0.292 | t-statistics = -1.38, p = 0.170 | t-statistics = -1.04, p = 0.300 |
| COMMS vs. PUN | t-statistics = -4.42, p = 0.000 | t-statistics = -3.31, p = 0.001 | t-statistics = -4.48, p = 0.000 | t-statistics = -4.26, p = 0.000 |
| COMMS vs. PUN&COMMS | t-statistics = -5.59, p = 0.000 | t-statistics = -4.05, p = 0.000 | t-statistics = -5.12, p = 0.000 | t-statistics = -5.63, p = 0.000 |
| BASE vs. PUN | F(1, 128) = 5.24, p = 0.0238 | F(1, 230) = 6.10, p = 0.0143 | F(1, 128) = 16.02, p = 0.0001 | F(1, 230) = 14.48, p = 0.0002 |
| BASE vs. PUN&COMMS | F(1, 128) = 12.75, p = 0.0005 | F(1, 230) = 11.23, p = 0.0009 | F(1, 128) = 24.16, p = 0.0000 | F(1, 230) = 30.95, p = 0.0000 |
| PUN vs. PUN&COMMS | F(1, 128) = 1.88 p = 0.1729 | F(1, 230) = 1.04, p = 0.3098 | F(1, 128) = 0.80, p = 0.3722 | F(1, 230) = 4.04, p = 0.0455 |

Note: The dependent variable is the contribution in the initial period (left two columns) or averaged over the five periods (right two columns). Robust standard errors are given in parentheses. COMMS is the default treatment. *, **, and *** denote 10%, 5%, and 1% significance levels (p-values are two-sided).

Table C.5: Logit model regressions: Individual contributions of 0 ECU

| | Model 1 | Model 2 | Model 3 | Model 4 |
|---|----------------------|----------------------|----------------------|----------------------|
| | Period 1: Gen 1 | Period 1: All gen | All Periods: Gen 1 | All Periods: All gen |
| BASE | -1.204** (0.552) | -0.511 (0.472) | 0.534 (0.618) | 0.420 (0.523) |
| PUN | -2.590*** (0.680) | -1.746*** (0.516) | -1.580*** (0.574) | -1.121** (0.457) |
| PUN&COMM | -4.066*** (1.103) | -2.326*** (0.527) | -2.140*** (0.604) | -2.404*** (0.469) |
| Constant | 0.511 (0.423) | -0.182 (0.350) | 0.887** (0.451) | 0.833** (0.380) |
| Observations | 132 | 234 | 132 | 234 |
| Pseudo R-squared | 0.235 | 0.125 | 0.189 | 0.189 |
| Pairwise comparisons of coefficients (Chi2 (1)) | | | | |
| COMMS vs. BASE | 4.75, p = 0.0293 | 1.17, p = 0.2796 | 0.75, p = 0.3875 | 0.65, p = 0.4218 |
| COMMS vs. PUN | 14.51, p = 0.0001 | 11.43, p = 0.0007 | 7.59, p = 0.0059 | 6.00, p = 0.0143 |
| COMMS vs. PUN&COMMS | 13.60, p = 0.0002 | 19.47, p = 0.0000 | 12.54, p = 0.0004 | 26.28, p = 0.000 |
| BASE vs. PUN | 4.69, p = 0.0303 | 6.24, p = 0.0125 | 14.68, p = 0.0001 | 12.22, p = 0.0005 |
| BASE vs. PUN&COMMS | 7.05, p = 0.0079 | 12.89, p = 0.003 | 20.99, p = 0.000 | 38.91, p = 0.000 |
| PUN vs. PUN&COMMS | 1.65, p = 0.1989 | 1.13, p = 0.2883 | 1.09, p = 0.2970 | 11.69, p = 0.0006 |

Note: The dependent variable is the contribution in the initial period (left two columns) or averaged over the five periods (right two columns). Robust standard errors are given in parentheses. COMMS is the default treatment. *, **, and *** denote 10%, 5%, and 1% significance levels (p-values are two-sided).

Why are participants punished?

Table C.6: Linear probability model regressions: Punishment received

| | Model 1 | Model 2 | Model 3 |
|---|--------------------|--------------------|--------------------|
| PUN&COMMS | -0.12*** (0.02) | -0.27*** (0.06) | -0.22*** (0.06) |
| Deviation from the others' average contribution | 0.03*** (0.01) | 0.05*** (0.01) | 0.04*** (0.01) |
| PUN&COMMS x Deviation from the others' average contribution | - | -0.06*** (0.01) | -0.06*** (0.01) |
| Average others' contribution | -0.01* (0.01) | -0.03*** (0.01) | 0.01 (0.02) |
| PUN&COMMS x Average others' contribution | - | 0.04*** (0.01) | 0.04* (0.02) |
| Contribute at least 15 in the period | - | - | -0.21*** (0.06) |
| PUN&COMMS x Contribute at least 15 in the period | - | - | -0.03 (0.09) |
| Constant | 0.23*** (0.05) | 0.31*** (0.06) | 0.23*** (0.06) |
| Period FE | Yes | Yes | Yes |
| Generation FE | Yes | Yes | Yes |
| Observations | 780 | 780 | 780 |
| R-squared | 0.12 | 0.16 | 0.20 |
| Pairwise tests | | | |
| Model 2: | | | |
| Deviation from others' average contribution + PUN&COMMS x Deviation from others' = 0 | | | |
| F(1, 768) = 0.08, p=0.7710 | | | |
| Average others' contribution + PUN&COMMS x Average others' contribution | | | |
| F(1, 768) = 1.67, p=0.1972 | | | |
| Model 3: | | | |
| Deviation from others' average contribution + PUN&COMMS x Deviation from others' = 0 | | | |
| F(1, 766) = 2.53, p=0.1123 | | | |
| Average others' contribution + PUN&COMMS x Average others' contribution | | | |
| F(1, 766) = 13.66***, P= 0.0002 | | | |
| Contribute at least 15 in the period + PUN&COMMS x Contribute at least 15 in the period | | | |
| F(1, 766) = 13.35***, p=0.0003 | | | |

Note: The dependent variable is the punishment received. Robust standard errors are given in parentheses. COMMS is the default treatment. *, **, and *** denote 10%, 5%, and 1% significance levels (p-values are two-sided).

Using a linear probability model in Table C.6, we find that participants were punished significantly less in PUN&COMMS than in PUN (17% of decisions were punished in PUN, whereas only 5% in PUN&COMMS). Our results remain similar using a random effect panel model instead (shown in Table C.7).³ The higher the average other members' contribution, the higher the punishment in PUN&COMMS but not in PUN. Finally, there is significantly less punishment in both treatments in periods with generations contributing 15 ECU or more (i.e., the focal contribution for cooperation). These latter three results are robust with respect to the estimation method (attained in both Tables C.6 and C.7). In PUN, only 24% of all members contribute exactly as many ECU as the average of the other two members compared to 78% in PUN&COMMS. Thus, it

³Note that the coefficient of PUN&COMMS turns insignificant in Model 3 of Table C.7, but that may be because, in almost all periods of this treatment, members of a generation contribute 15 or more.

appears that PUN&COMMS attains high levels of sustained generations without executing much punishment.

Table C.7: Random-effects panel data linear probability model regressions: Punishment received

| | Model 1 | Model 2 | Model 3 |
|---|--------------------|-------------------|--------------------|
| PUN & COMS | -0.12** (0.06) | -0.21* (0.12) | -0.18 (0.12) |
| Deviation from the others' average contribution | 0.02*** (0.00) | 0.04*** (0.00) | 0.03*** (0.00) |
| PUN&COMMS x Deviation from the others' average contribution | - | -0.04*** | -0.05*** |
| Average others' contribution | -0.01*** (0.00) | (0.01) -0.02* | (0.01) 0.01 |
| PUN&COMMS x Average other's contribution | | 0.02 (0.02) | 0.03** (0.01) |
| Contribute at least 15 in the period | | | -0.15*** (0.05) |
| PUN&COMMS x Contribute at least 15 in the period | | | -0.03 (0.02) |
| Constant | 0.23*** (0.06) | 0.28*** (0.08) | 0.22** (0.09) |
| Period FE | YES | YES | YES |
| Generation FE | YES | YES | YES |
| Observations | 780 | 780 | 780 |
| R-squared | 0.12 | 0.15 | 0.21 |

Pairwise tests

Model 2:

Deviation from others' average contribution + PUN&COMMS x Deviation from others' = 0
Chi-squared(1) = 0.05 , p=0.8266

Average others' contribution + PUN&COMMS x Average others' contribution
Chi-squared(1) = 0.01 , p=0.9116

Model 3:

Deviation from others' average contribution + PUN&COMMS x Deviation from others' = 0
Chi-squared(1) = 1.15 , p=0.2833

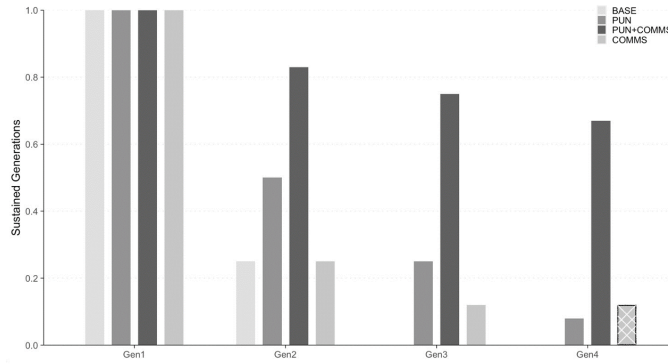
Average others' contribution + PUN&COMMS x Average others' contribution
Chi-squared(1) = 7.45*** , p=0.0064

Contribute at least 15 in the period + PUN&COMMS x Contribute at least 15 in the period
Chi-squared(1) = 6.59** , p=0.0102

Note: The dependent variable is the punishment received. Robust standard errors are given in parentheses. COMMS is the default treatment. *, **, and *** denote 10%, 5%, and 1% significance levels. p-values are two-sided.

D Robustness analysis regarding the lost observation

In COMMS, there was one cohort that reached Gen 3, but at this point, a member of this generation left the experiment. Thus, we had to stop the experiment at Gen 3 without knowing what would have happened if the participant had stayed. In a sense, leaving the experiment may be interpreted as a form of free riding, so it may be reasonable to assume that the cohort reached Gen 3. In the following, we study whether this assumption affects the results.



Note: In BASE, 12 cohorts started in Gen 1, and 3, 0, and 0 cohorts continued to Gen 2, Gen 3, and Gen 4. In PUN, 12 cohorts started in Gen 1, and 6, 3, and 1 cohorts continued to Gen 2, Gen 3, and Gen 4. In PUN+COMMS, 12 cohorts started in Gen 1, and 10, 9, and 8 continued to Gen 2, Gen 3, and Gen 4. In COMMS, 8 cohorts started in Gen 1, and 2 and 1 cohorts reached Gen 2 and 3. Then, one participant in Gen 3 dropped off the experiment, and we had to end the experiment for this cohort while the cohort could have reached Gen 4. This is marked in the figure.

Figure D.1: Sustained generations across treatments considering the lost observation in COMMS

When inspecting the sequence of maximal generations reached, Table D.1 shows that when considering the lost observation in the COMMS treatment, the average maximal generation slightly increases from 1.375 to 1.5.

Table D.1: Descriptive statistics: Occurrence of sustained generations

| | Average | Median | SD | Min | Max | Observations |
|-----------|---------|--------|------|-----|-----|--------------|
| BASE | 1.25 | 1 | 0.45 | 1 | 2 | 12 |
| PUN | 1.833 | 1.5 | 1.03 | 1 | 4 | 12 |
| PUN+COMMS | 3.25 | 4 | 1.22 | 1 | 4 | 12 |
| COMMS | 1.375 | 1 | 0.74 | 1 | 3 | 8 |
| COMMS* | 1.5 | 1 | 1.07 | 1 | 4 | 8 |

Note: *Occurrence of sustained generations* is a variable ranges between 1-4, indicating the number of sustained generations. COMMS does not consider that the particular cohort reached Gen 4, while COMMS* considers it.

Note that the addition of one successful generation in one COMMS cohort (due to the dropout) does not change the results of the robust rank order tests or the Chi-squared tests. Still, there are significant differences at less than 1% level between PUN&COMMS and any of the other treatments (see Tables D.2 and D.3).

Table D.2: Occurrence of sustained generations in the different treatments, based on Table B.1

| | Observations | U - statistic |
|-----------------------|--------------|---------------|
| BASE vs. COMMS: | 12 vs. 8 | -0.106 |
| BASE vs. COMMS*: | 12 vs. 8 | -0.106 |
| BASE vs. PUN: | 12 vs. 12 | -1.299* |
| BASE vs. PUN&COMMS: | 12 vs. 12 | -3.774*** |
| COMMS vs. PUN: | 8 vs. 12 | -0.974 |
| COMMS vs. PUN&COMMS: | 8 vs. 12 | -3.321*** |
| COMMS* vs. PUN: | 8 vs. 12 | -0.828 |
| COMMS* vs. PUN&COMMS: | 8 vs. 12 | -2.585** |
| PUN vs. PUN&COMMS: | 12 vs. 12 | -2.743** |

Note: Two-sample Fligner-Policello robust rank order test. The dependent variable, *Occurrence of sustained generations*, ranges between 1-4, indicating the number of sustained generations. *, **, and *** denote 10%, 5%, and 1% significance levels. Two-sided small-sample using (Feltovich, 2005). COMMS does not consider that the particular cohort reached Gen 4, while COMMS* considers it.

Table D.3: Success rate in sustaining generations across all possible generations, based on Table B.2

| | Observations | $\chi^2(1)$ -statistic | P-value (two-sided) |
|-----------------------|--------------|------------------------|---------------------|
| BASE vs. COMMS: | 36 vs. 24 | 0.2778 | 0.598 |
| BASE vs. COMMS*: | 36 vs. 24 | 0.9704 | 0.325 |
| BASE vs. PUN: | 36 vs. 36 | 4.5997** | 0.032 |
| BASE vs. PUN&COMMS: | 36 vs. 36 | 32.9143*** | 0.000 |
| COMMS vs. PUN: | 24 vs. 36 | 1.9804 | 0.159 |
| COMMS vs. PUN&COMMS: | 24 vs. 36 | 22.5000*** | 0.000 |
| COMMS* vs. PUN: | 24 vs. 36 | 0.9938 | 0.319 |
| COMMS* vs. PUN&COMMS: | 24 vs. 36 | 19.6218*** | 0.000 |
| PUN vs. PUN&COMMS: | 36 vs. 36 | 16.0680*** | 0.000 |

Note: Pairwise Chi-squared tests. In creating the dependent variable, *Success rate in sustaining generations*, We assume that any cohort has 3 generations. For example, if a cohort reached Gen 4, the success is (1,1,1), whereas if a cohort failed in Gen 1, the data are recorded as (0,0,0). Finally, *, **, and *** denote 10%, 5%, and 1% significance levels(two sided). COMMS does not consider that the particular cohort reached Gen 4, while COMMS* considers it.

E Bootstrap simulations

Because our experiment requires an independent cohort of 12 participants, we have a limited number of cohorts (12 in BASE, PUN, and PUN&COMMS and

8 in COMMS). The number of realizations (affected by the random assignment of participants into generations) is small compared to all possible combinations. Assigning participants to generations and placing them at a specific position of the generational sequence can further impact the number of sustained generations. Therefore, to test the robustness of our main result, we conducted a bootstrap simulation similar to [Hauser et al. \(2014\)](#) and [Lohse and Waichman \(2020\)](#). We simulated 10,000 randomly generated generational sequences for each treatment (in total, 40,000). In each cohort, we randomly sample (with replacement) a series of participants' contribution decisions for each generation, calculate the total amount contributed to the group account, and observe whether the generation was sustained. In particular, our procedure works as follows (for each treatment):

1. For the first generation, we consider only participants of Gen 1 in the respective treatment. We randomly sample (with replacement) three of their contribution decisions in the first period. Based on the contribution decisions, we calculate the total amount contributed to the group account by multiplying the sum of the three contribution decisions by 5 (due to the 5 periods) and observe whether the sum is above the threshold.
2. For the second and third generations, we consider all generations in the respective treatment. We randomly sample (with replacement) three of their contribution decisions in the first period. Based on the contribution decisions, we calculate the total amount contributed to the group account by multiplying the sum of the three contribution decisions by 5 (due to the 5 periods) and learn whether the sum is above the threshold.
3. Finally, we determine the point at which the generational sequence would break down (if any) and start again at step 1.

Using this procedure, we simulated 10,000 cohorts of 4 generations. For BASE, COMMS, and PUN, the simulated mean value of sustained generations is very close to the results of our experiment (BASE: 1.4 in the simulation vs. 1.3 in the experiment, COMMS: 1.3 vs. 1.4, and PUN: 1.8 vs. 1.8). Even though, PUN&COMMS still reaches considerably higher sustained generations than the other treatments, the mean of the simulation is considerably lower than the mean in our experiment (2.6 vs. 3.3).

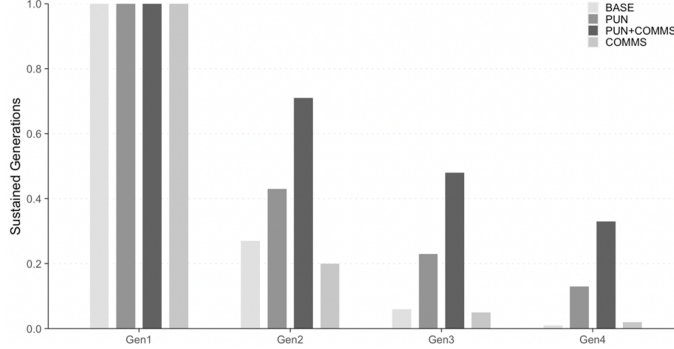


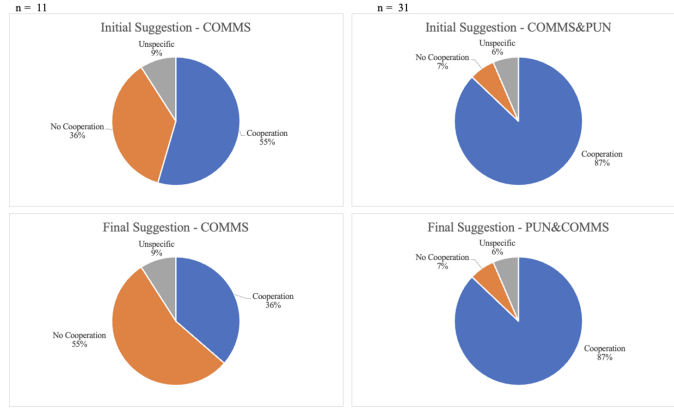
Figure E.1: Simulation result

A reason for the low performance of the simulation in PUNSS&COMMS is that while the simulation approach may be particularly useful for BASE and PUN, it is not the case for communication. In COMMS and PUN&COMMS, there is a communication phase before the first period, and thus, randomly sampling 3 participants from different generations will likely not reflect the effect of communication, which allows all three members of the generation to draw inferences on generation composition (whether to cooperate or to defect fully). Indeed, in COMMS and PUN&COMMS, first-period members coordinated well to contribute 25 each (when they wanted to reach the threshold) or 0 (when they did not want to reach the threshold). Another limitation of the simulation for our game is that we can only use the first-period (i.e., independent) contribution decisions for our simulation. If most cooperation or coordination happens at later periods, it would not be reflected in the simulation. Yet, the simulation results in Figure E.1 demonstrate that first-period contributions provide a rather good prediction of success.

F Analysis of communication

We analyzed the chat messages to learn how communication was used with and without punishment. Four annotators received annotation guidelines (see Section F.2) to evaluate the content of the chat messages. The annotators received one log file for each generation with all chat messages written in all five periods (see an example below). For each logfile, they answered two questions about whether there was a suggestion “to cooperate,” “not to cooperate,” or “no suggestion,” and three questions regarding punishment (only for PUN&COMMS). In 71% of all cases, all four annotators came to the same conclusion. In 26% of the cases, three annotators came to the same conclusion. In this case, we applied a majority rule (i.e., we selected the interpretation of the three annotators). Finally, in 3% of the cases, only two annotators came to the same conclusion. We marked those chats as “unspecified.”

F.1 Results of the communication analysis



Note: The top panel answers question 1: What was the initial suggestion in the chat? (“Initial Suggestion”). The bottom panel answers question 2: What was the final suggestion at the end of the chat? (“Final Suggestion”).

Figure F.1: Communication analysis

The annotators’ conclusions about the content of the chat messages are as follows: In PUN&COMMS, the first suggestion in the chat was more often to cooperate than in COMMS. This could indicate that in the presence of punishment, payoff-maximizing participants are more hesitant to reveal their type early in the chat. In COMMS, the annotators recognized that some payoff-maximizing participants used the chat to reveal their strategy. Without punishment, participants who want to cooperate do not have any leverage against them (and in fact should not cooperate if they believe that the threshold is not met), leading to an increase in uncooperative messages throughout the chat (see the final suggestions in the bottom panel of Figure F.1). Contrary, the annotators did not find a change in the generation’s intent to cooperate in PUN&COMMS. This might be because payoff-maximizing participants were more hesitant to influence the discussions (as they might have feared being punished).

Next, in PUN&COMMS, the annotators found only very few occasions (6%) in which participants explicitly threatened other participants with punishment. Instead, on many more occasions (19%), participants agreed not to punish each other if they stuck to what the decision makers agreed on. In PUN&COMMS, the participants rather used the chat to agree on the focal point of cooperation (5 ECU per participant per period). It seems that the presence of punishment leads to more cooperative behavior as communication makes the presence of participants who may punish sufficiently prominent to force payoff-maximizing participants into cooperation.

F.2 Annotation guidelines

The following guidelines were given to the annotators:

Guidelines

Your task is to read through the chat messages and answer a few questions for every group. The answer to each question might not always be clear, and even with the same annotation guideline, annotators can come to different answers. To counterbalance these inconsistencies, the chat messages are annotated by 4 annotators. After the annotation process, in case of disagreement, we use the majority opinion to determine the outcome.

The experiment

A cohort of 12 participants is divided into four generations (labeled as Gen 1, Gen 2, Gen 3, and Gen 4). Each generation consists of three participants. Members of Gen 1 always actively participate. However, the members of subsequent generations only participate depending on the decisions of the previous generations. For each participating generation, there are five decision periods.

For each decision period, each participant receives an endowment of 10 Experimental Currency Units (i.e., in the total of five periods, each member makes decisions regarding an overall amount of 50 ECU). In each decision period, each participant decides how much of the 10 ECU to keep and how many ECU to assign to a group account. The amount that a participant keeps will be paid to this participant at the end of the experiment. The ECU in the group account will be added up. After each period, each group member will be able to see how many ECU the two other group members have assigned individually to the group account, as well as the sum of the amounts assigned to this account. If after the five decision periods, the sum assigned to the group account by the three group members is 75 ECU or more, the experiment continues, and the next group starts. If the sum is less than 75 ECU, the experiment ends. In this case, all participants of the next groups (please note that we use the terms group and generation interchangeably) do not earn money from the experiment.

In particular, if Gen 1 assigned a total of at least 75 ECU to the group account, then the experiment continues to Gen 2. If Gen 2 assigned at least 75 ECU to the group account, then the experiment continues to Gen 3. If Gen 3 assigned at least 75 ECU to the group account, then the experiment continues to Gen 4. Finally, if the experiment reaches Gen 4, each member of Gen 4 earns 25 ECU, and the experiment ends. If it does not reach Gen 4, the members will earn nothing from the experiment. In each decision period, after finding out how much the other members of the generation have assigned to the group account, participants have the possibility to penalize members of their generation. They can assign to each of the generation members 0, 1, or at most 2 negative points. Each negative point costs 1 ECU. Each negative point that they assign is multiplied by three and reduces the remaining ECU of the penalized member (i.e., the group member who received the negative

points). E.g., if a participant assigns one negative point, the remaining ECU of the penalized member is reduced by 3 ECU. Before each decision period, the participants had the opportunity to communicate via chat messages which represents the basis of this analysis.

The chat messages

You receive a folder with 14 excel spreadsheets. Please note:

- Each spreadsheet represents one session.
- Each session can contain up to two cohorts.
- Each cohort consists of up to four generations/groups. If a generation fails to reach the threshold, the subsequent generations are not included.
- Each generation/group consists of three participants playing five periods.

You can find all messages of one generation in one sheet within the spreadsheet, please see an example in Figure F.2. Please note that some participants did not use the chat. Therefore, their ID might not appear in the respective column (“nickname”). In each sheet, you find a set of questions (not all of them cover questions on punishment; see Figure F.3). Please use the dropdowns to answer them. Please make sure you answered all questions on all (!) sheets.

| Session | id_p_und | participant | channel | nickname | chat | time | Gruppe | Generations |
|----------|----------|-------------|---------|----------|------|------|--------|-------------|
| Period 1 | | | | | | | | |
| Period 2 | | | | | | | | |
| Period 3 | | | | | | | | |
| Period 4 | | | | | | | | |
| Period 5 | | | | | | | | |

Figure F.2: Example of one generation’s chat

Questions

| | |
|--|--|
| <p>Bitte beantworten Sie die folgenden Fragen:</p> <p>1. Was war der erste Vorschlag im Chat?</p> <p>2. Was war der aktuelle Vorschlag am Ende des Chats?</p> <p>3. Wurde über die Negativpunkte im Allgemeinen gesprochen?</p> <p>Bitte beantworten Sie die folgenden Fragen nur, wenn sie Frage 3 mit "Ja" beantwortet haben.</p> <p>3a Gab es innerhalb der Diskussion eine Drohung Negativpunkte zu verteilen?</p> <p>3b Gab es innerhalb der Diskussion eine Vereinbarung keine Negativpunkte zu verteilen?</p> | <p>Session 4y11nwa3</p> <p>Gruppe 2</p> <p>Generation G3</p> |
|--|--|

Figure F.3: Questions

Question 1 (“What was the initial suggestion in the chat?”): You can choose between “To cooperate,” “Not to cooperate,” and “Undefined.” Please only consider the very first suggestion of any of the participants. Please see an example in Figure F.4. We define cooperation as the will to reach the threshold of 75 ECU.

The screenshot shows a chat window with a list of messages from participants. The first message is highlighted with a yellow box labeled "First Suggestion". To the right, there is a survey titled "Bitte beantworten Sie die folgenden Fragen:" with two questions and two buttons: "Zu kooperieren" and "Nicht zu kooperieren".

| nickname | body |
|--------------|--|
| Teilnehmer 1 | hallo |
| Teilnehmer 3 | Ihr pusst was auf! Gruppenkonto oder? |
| Teilnehmer 2 | muh |
| Teilnehmer 2 | seiner! Verliere ich dadurch nicht nur Geld? |
| Teilnehmer 3 | Mmh okay, also ich würde schon was draufpacken, ist gerechter |
| Teilnehmer 1 | wer meinten ja nicht sozial sein |
| Teilnehmer 3 | aber wenn ihr es nicht draufschafft, kann ich es auch lassen, dann litz ja eh egal |
| Teilnehmer 1 | hat es Nachteile nichts draufpacken? |
| Teilnehmer 2 | Was haben wir davon, wenn es durch alle Gruppen durchgeht? |
| Teilnehmer 3 | ich glaube keine Nachteile |
| Teilnehmer 1 | dann lass nichts draufpacken |

First Suggestion

Bitte beantworten Sie die folgenden Fragen:

1. Was war der erste Vorschlag im Chat?

2. Was war der aktuelle Vorschlag am Ende des Chats?

Zu kooperieren

Nicht zu kooperieren

Figure F.4: Example for Questions 1 and 2

Question 2 (“What was the final suggestion in the chat?”): You can choose between “To cooperate,” “Not to cooperate,” and “Undefined.” Please only consider the very last suggestion of any of the participants. Please see an example in Figure F.4. We define cooperation as the will to reach the threshold of 75 ECU.

The screenshot shows a chat window with a list of messages from participants. The last message is highlighted with a yellow box labeled "Final Suggestion". To the right, there is a survey titled "Bitte beantworten Sie die folgenden Fragen:" with three questions and two buttons: "Ja" and "Nein".

| nickname | body |
|--------------|--|
| Teilnehmer 6 | Es reicht, wenn jeder je Runde 5 gibt |
| Teilnehmer 5 | Gesau, jeder immer 5 |
| Teilnehmer 5 | Dachte, gleich alles rein und die letzten halt 0 |
| Teilnehmer 5 | aber na gut |
| Teilnehmer 5 | kommt aufs Gleiche hinaus :) |
| Teilnehmer 6 | Bist keine Negativpunkte |
| Teilnehmer 5 | dann gebe ich mychste 0 und danach 5 |
| Teilnehmer 6 | Easy! |

Final Suggestion

Bitte beantworten Sie die folgenden Fragen:

3. Wurde über die Negativpunkte im Allgemeinen gesprochen?

Bitte beantworten Sie die folgenden Fragen nur, wenn sie Frage 3 mit "ja" beantwortet haben.

3a. Gab es innerhalb der Diskussion eine Ordnung Negativpunkte zu verteilen?

3b. Gab es innerhalb der Diskussion eine Vereinbarung keine Negativpunkte zu verteilen?

Ja

Nein

Ja

Nein

Figure F.5: Example for Questions 3, 3a, and 3b

Question 3 (“Did the participants discuss the negative points in general?”), 3a (“Did at least one participant threaten others to punish them?”), 3b (“Did at least one participant suggest not to punish each other?”): Please note that the questions are not relevant for all sessions; i.e., you will find sheets without those questions. If you answer question 3 with no, you can leave questions 3a and 3b blank. Please see an example in Figure F.5.

Please do not hesitate to reach out to stenzel@economics.uni-kiel.de in case of any questions. Thank you very much!

References

- Bochet, O., Page, T., and Putterman, L. (2006). Communication and punishment in voluntary contribution experiments. *Journal of Economic Behavior & Organization*, 60(1):11–26.
- Bock, O., Baetge, I., and Nicklisch, A. (2014). hroot: Hamburg registration and organization online tool. *European Economic Review*, 71:117–120.
- Chen, D. L., Schonger, M., and Wickens, C. (2016). oTree—an open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, 9:88–97.
- Feltovich, N. (2005). Critical values for the robust rank-order test. *Communications in Statistics - Simulation and Computation*, 34(3):525–547.
- Hauser, O. P., Rand, D. G., Peysakhovich, A., and Nowak, M. A. (2014). Co-operating with the future. *Nature*, 511(7508):220–223.
- Lohse, J. and Waichman, I. (2020). The effects of contemporaneous peer punishment on cooperation with the future. *Nature Communications*, 11(1):1815.
- Milinski, M., Sommerfeld, R. D., Krambeck, H.-J., Reed, F. A., and Marotzke, J. (2008). The collective-risk social dilemma and the prevention of simulated dangerous climate change. *Proceedings of the National Academy of Sciences*, 105(7):2291–2294.
- Schelling, T. (1960). *The Strategy of Conflict*. Harvard University Press, Cambridge, MA.
- Tavoni, A., Dannenberg, A., Kallis, G., and Löschel, A. (2011). Inequality, communication, and the avoidance of disastrous climate change in a public goods game. *Proceedings of the National Academy of Sciences*, 108(29):11825–11829.

G Translated instructions (computer screens)

This section includes the translated instructions for the PUN&COMMS treatment. The PUN (COMMS) treatment instructions are identical, excluding the COMMS (PUN) parts. The baseline instructions are identical, excluding both parts for COMMS and PUN. Each subsection represents a single screen presented to the participants.

Welcome

Dear participants,

Thank you for participating in our experiment. The experiment is expected to last 60 minutes. The experiment is carried out in strict compliance with the code of conduct for economic experiments of the University of Kiel. This means, among other things, that all information is provided to you truthfully. You can end the experiment at any time without giving a reason. Your data will be treated as strictly confidential and will be processed and published anonymously by the researchers involved. In the following, for reasons of better readability, only the female form will be used, although this explicitly refers to people of any gender.

All participants receive the same instructions. Please read the instructions carefully.

Please mute your microphone. If you have any questions, write to us using the private chat feature in Zoom.

Each participant receives €5.00 as a participation fee, regardless of the course of the experiment. Depending on your decisions and the decisions of the other participants during the experiment, you can earn more money.

During the experiment we use Experimental Currency Units (ECU) instead of euro as currency. The amount you get for the decision tasks is calculated and displayed in ECU. At the end of the experiment, the total amount will be converted into euro and (in addition to the €5.00 participation fee) transferred to your specified bank account within 2-3 days after the experiment. At the end of the experiment, we will ask you for your account information.

The exchange rate between ECU and euro is 2.50 : 1. This means, for every 2.50 ECU you earn, you receive €1.00.

Group Assignment and Experiment Procedure

In this experiment, every 12 participants will be randomly and anonymously divided into four groups, each group consists of three participants. The four

groups will be labeled during the experiment as Group 1 (with members 1, 2, and 3), Group 2 (with members 4, 5, and 6), Group 3 (with members 7, 8, and 9), and Group 4 (with members 10, 11, and 12). A random mechanism will determine if you will be assigned to Group 1, 2, 3, or 4. Once the computer program starts, you will see on your computer screen if you are assigned to Group 1, 2, 3, or 4.

In this experiment, members of Group 1 always actively participate (i.e., taking decisions). However, the members of subsequent groups only participate depending on the decisions of the previous groups. I.e., Group 2's participation depends on the decisions of Group 1, and similarly, Group 3's participation depends on the decisions of Group 1 and Group 2. Finally, Group 4's participation depends on the decisions of Group 1, Group 2, and Group 3.

The Order of Decisions

For each group, there are five decision rounds. You take two decisions per round. All members of the group simultaneously make their decisions. For each decision round, each participant receives an endowment of 10 ECU (i.e., in the total of five rounds, each member makes decisions regarding an overall amount of 50 ECU).

Before each decision round, the 3 group members can chat with each other via a chat window on your computer screen. In the first two rounds, you have 3 minutes, and in the rounds 3-5, you have 1.5 minutes per round to communicate. You can only talk with the two other members of your group. You are not allowed to identify yourself or other group members in this chat.

In each decision round, each participant decides how much of the 10 ECU to keep and how much ECU to assign to a group account. The amount that a participant keeps will be paid to this participant at the end of the experiment. The ECU in the group account will be added up.

After each round, each group member will be able to see how many ECU the two other group members have assigned individually to the group account, as well as the sum of the amounts assigned to this account.

If after the five decision rounds, the sum assigned to the group account by the three group members is 75 ECU or more, the experiment continues, and the next group starts. If the sum is less than 75 ECU, the experiment ends. In this case, all participants of the next groups do not earn money from the experiment.

In particular, if Group 1 assigned a total of at least 75 ECU to the group account, then the experiment continues to Group 2. If Group 2 assigned at least 75 ECU to the group account, then the experiment continues to Group 3. If Group 3 assigned at least 75 ECU to the group account, then the experiment

continues to Group 3. Finally, if the experiment reaches Group 4, each member of Group 4 earns 25 ECU, and the experiment ends. If it does not reach Group 4, the members will earn nothing from the experiment.

Note that besides the earnings from the experiment that depend on your decisions and the decisions of previous groups, every participant is entitled to €5.00 as a participation fee, irrespective of the course of the experiment.

Negative Points

After being informed in each round of decisions how much the two other participants in your group have sent to the group account, you make the second decision. You can attribute negative points to the two other participants in your group. You can assign to each of your group members 0, 1, or at most 2 negative points. Each negative point that you assign costs 1 ECU. Each negative point that you assign is multiplied by three and reduces the remaining ECU of the penalized member. You have three possibilities:

- You assign 0 negative points (no costs for you and the remaining ECU of the penalized member is reduced by 0 ECU).
- You assign 1 negative point (costs you 1 ECU and the remaining ECU of the penalized member is reduced by 3 ECU).
- You assign 2 negative points (costs you 2 ECU and the remaining ECU of the penalized member is reduced by 6 ECU).

After your decision, you can see how many negative points you assigned and how many negative points you have received. All decisions are anonymous: You cannot find out which participant has assigned negative points to you, and others cannot find out to which participant you have assigned negative points.

Earning per Round and Total Earning

In case the previous group contributed enough ECUs to the group account, your group was reached, and you took decisions; then, your payoff for each round will be as follows:

- Endowment for this round (10 ECU),
- minus the assigned amount to the group account,
- minus assigned negative points,
- minus received negative points times 3.

Please note that your payoff for one round can be negative. Your payoff is summed up for the five decision rounds. In cases where your summed payoff is negative, it will be set to 0 (and in any case, you will earn the participation fee

of €5.00). In case your group was not reached, and you did not take decisions: You will not earn money from the experiment. However, you, as everybody else, are entitled to the €5.00 participation fee.

Test Questions

Please consider the following cases and answer the questions. At the end of the page, you will also find a summary of the instructions.

Case 1: After the five decision rounds, Participant 1 in Group 1 assigned a total of 30 ECU to the group account. Moreover, after the five decision rounds, the total amount in the group account of Group 1 is 75 ECU.

- What is the total payoff for Participant 1 (in ECU): [20 ECU]
- Does the experiment continue to Group 2? [yes]

Case 2: After the five decision rounds, Participant 1 in Group 1 assigned a total of 20 ECU to the group account. Moreover, after the five decision rounds, the total amount in the group account of Group 1 is 70 ECU.

- What is the total payoff for Participant 1 (in ECU): [30 ECU]
- Does the experiment continue to Group 2? [no]
- What is the total payoff for members in Group 2 (in ECU): [0 ECU]

Case 3: Participant 2 in Group 2 assigned 0 negative points to Participant 1 and 2 negative points to Participant 3.

- By how many ECU is the payoff of Participant 2 (the sender) reduced? [2]
- By how many ECU is the payoff of Participant 1 reduced? [0]
- By how many ECU is the payoff of Participant 3 reduced? [6]

The experiment will start once you press “continue.”

Group Assignment

You are Participant [x] and are assigned to Group [y]. {For participants in Groups 2, 3, and 4:} Group 1 will start. Please wait until they make their decisions. This can take up to 15 minutes.

Chat

Remaining time: [Time]

You are Participant [x] in Group [y]. This is Round [z] of 5. You can now chat with your other 2 group members for 3 minutes / 1.5 minutes. You can leave the chat already earlier. You are not allowed to identify yourself or other group members in this chat. In the next round, you can chat with your other 2 group members again. If you want to leave the chat earlier, please click “next.” Please note that you cannot return to the chat in this round.

Group Account

You are Participant [x] in Group [y]. This is Round [z] of 5.

For this task, you and two additional group members have received an endowment of 10 ECU. On this screen, you can decide how many (0-10) of your ECU you want to assign to the group account. All ECU that would be assigned to this group account will be added up. If after five rounds the sum is equal or larger to 75 ECU, the next group will start to play. If the sum accumulated in the group account after 5 rounds is less than 75 ECU, the experiment will end, and the next group will not make decisions (and will not receive an additional endowment).

Please state how many ECU you want to send to the group account: [...]

Contributions to the Group Account

You are Participant [x] in Group [y]. This is Round [z] of 5. Here, you are informed about the decisions of the participants in your group.

This Round:

- Your contribution: ...
- Participant [1]: ...
- Participant [2]: ...

Up to this round:

- Your contribution: ...
- Participant [1]: ...
- Participant [2]: ...

You and the other members of your group contributed $[\sum c_i]$ ECU in total up to this point. *{In the last round:}* There are more than 75 ECU in the group account and, therefore, the next group starts. / There are not more than 75 ECU in the group account and, as a consequence, the experiment ends.

Negative Points

You are Participant [x] in Group [y]. This is Round [z] of 5. On this screen, you can take a decision if you want to assign negative points to another member. A negative point costs you 1 ECU and reduces the earnings of the other participant by 3 ECU. You can assign between 0 and 2 negative points to each participant in each round.

Round Payoff

You are Participant [x] in Group [y]. This is Round [z] of 5.

- Your endowment: ...
- Your assignment to the group account: ...
- Your assigned negative points: ...
- Negative points received: ...
- **Your round payoff:** ...

Total Payoff

You are Participant [x] in Group [y]. This is Round [z] of 5. Your payoffs per round:

- Round 1: ...
- Round 2: ...
- Round 3: ...
- Round 4: ...
- Round 5: ...
- **Total:** ...