Abstract

We explore the relationship between individuals’ disposition to cooperate and their inclination to engage in peer punishment as well as their relative importance for mitigating social dilemmas. Using a modified strategy-method approach we identify individual punishment patterns and link them with individual cooperation patterns. Classifying \( N = 628 \) subjects along these two dimensions documents that cooperation and punishment patterns are aligned for most individuals. However, the data also reveal a sizable share of free-riders that punish pro-socially and conditional cooperators that do not engage in punishment. Analyzing the interplay between types in an additional experiment, we show that pro-social punishers are important for achieving cooperation. Incorporating information on punishment types explains large amounts of the between- and within-group variation in cooperation.

JEL-Classification: C9; D03

Keywords: strategy-method, punishment patterns, type classification, conditional cooperation, public-goods game

Acknowledgements: We would like to thank the editor, Tim Cason, two anonymous referees as well as Nick Bardsley, Christoph Engel, Louis Putterman, Christian Thöni and numerous seminar and workshop participants for helpful comments and suggestions. Financial support from DFG (Deutsche Forschungsgemeinschaft; grant 50130225) is gratefully acknowledged. The usual disclaimer applies.

† University of Marburg; ‡ University of Bonn; § Max Planck Institute for Research on Collective Goods; * Hertie School of Governance, Berlin.
1 Introduction

An extensive body of research documents cooperation among humans (e.g., Andreoni, 1988; Ledyard, 1994; Fischbacher and Gächter, 2010; Balliet et al., 2011; Chaudhuri, 2011, to name only a few), pointing out that cooperation problems can be mitigated by appropriate institutional settings (e.g., Ostrom et al., 1992; Kosfeld et al., 2009). Among these, the ubiquitous mechanism of peer punishment plays a prominent role in the literature (e.g., Fehr and Gächter, 2000, 2002; Carpenter, 2007; Reuben and Riedl, 2013). Even though peer punishment makes successful cooperation much more likely to occur, there are still groups who fail to use decentralized punishment in an effective and pro-social manner. This might be due to the fact that peer punishment constitutes a cooperation problem in itself (Yamagishi, 1986). A breakdown in cooperation that coincides with a failure of peer punishment could thus capture two sides of the same coin (see, e.g., Ones and Putterman, 2007; Peysakhovich et al., 2014). This conjecture raises two fundamental questions that we try to answer in this paper: Firstly, what is the relation between an individual’s disposition to cooperate (Fischbacher et al., 2001; Fischbacher and Gächter, 2010) and her individual inclination to engage in peer punishment? Secondly, if these two dispositions do not coincide, which of the two is relatively more important in achieving cooperative outcomes under peer punishment?

We study these questions employing a classical workhorse in the literature on cooperation and punishment: a linear public-goods game (VCM) with decentralized punishment (Fehr and Gächter, 2002). Subjects first make a contribution decision and can then assign costly punishment points that reduce the other group members’ payoffs. Within this prominent paradigm, we introduce a variant of the strategy-method at the punishment stage of the game that allows identifying heterogeneity in peer punishment at the individual level.

When making her punishment decisions, each subject is confronted with a random sequence of ‘scenarios’, i.e., combinations of others’ contributions. One of these scenarios corresponds to the other group members’ actual contribution decisions. All other scenarios are randomly drawn contributions that systematically cover relevant parts of the strategy space. Only the punishment decisions for the scenario with the actual contributions become payoff-relevant. As subjects do not know which scenario is the ‘relevant’ one, we have an incentive compatible strategy-method that induces exogenous variation in others’ contributions to consistently estimate individual peer punishment patterns in a one-shot game (see Bardsley, 2000, for a related approach eliciting cooperation patterns).\(^1\)

\(^1\)An alternative approach, based on a conventional strategy-method together with a strongly restricted choice set, is implemented by Cheung (2014) and Kamei (2014), who offer interesting complementary findings on cooperation and punishment patterns, respectively. Beyond method and sample size, the present paper also differs from these studies in that we analyze the link between cooperation and punishment types as well as the role of the different types for achieving cooperative outcomes in a repeated game.
Using this strategy-method to elicit punishment patterns reveals substantial heterogeneity between individuals. In our sample of $N = 628$ experimental participants two patterns dominate: Almost every second subject (47.1%) is classified as a pro-social punisher. Their individual punishment patterns are all significantly decreasing in the other’s contributions, i.e., they target their punishment towards those contributing nothing or little to the public good. The second-largest group (40.3%) are non-punishers (‘second-stage free-riders’), i.e., subjects that do not at all engage in peer punishment. Beyond these two dominant types, there is only a small fraction of subjects that displays either an unsystematic pattern or a pattern that is increasing in the other’s contribution (in the spirit of ‘anti-social punishment’; see, e.g., Herrmann et al., 2008). Moreover, we document that among pro-social punishment types, patterns are almost exclusively ‘self-centered’ around the own contribution level.

Linking individual punishment patterns to the corresponding individual dispositions to cooperate — which we obtain from a within-subject design using the measure of conditional cooperation introduced in Fischbacher et al. (2001) — yields a two-dimensional classification that reveals two behavioral archetypes. (i) For the majority of our subjects cooperation and punishment types are aligned: we find that 55% of conditional cooperators punish pro-socially and that 56% of free-riders are non-punishers. (ii) Consequently, this also implies that a significant share of subjects have individual punishment- and cooperation-patterns that are diverging: 35% of conditional cooperators are non-punishers and 32% of free-riders do engage in pro-social punishment.

The ability to identify these two behavioral archetypes — individuals whose cooperation and punishment patterns are either aligned or diverging — is a major benefit from combining our approach to classify punishment patterns at the individual level with the conditional cooperation-measure from Fischbacher et al. (2001). Moreover, as the individuals’ inclinations to cooperate and to punish are far from being perfectly correlated, we can assess their respective importance for mitigating a social dilemma in the presence of punishment opportunities. To do so, we use these individual type-classifications from two one-shot games to explain group outcomes in a third game: a finitely repeated public-goods game with peer punishment — both among stable groups where players interact repeatedly (partner design) and among steadily alternating groups where a group’s type composition changes over time (stranger design).

In both conditions, we observe that groups with more conditional cooperators achieve higher average contributions that are also more stable over time, than groups with fewer conditional cooperators. While these observations mirror previous findings that highlight the important role of conditional cooperators (e.g., Gächter and Thöni, 2005), we also obtain a similar picture with respect to the group members’ punishment types. In fact, variation in punishers’ types seems to be crucial in this richer environment: keeping constant the fraction of conditional cooperators, average contributions are significantly higher in groups that contain more pro-social punishers.
The presence of pro-social punishers induces higher contributions among both, subjects classified as free-riders and among conditional cooperators.

These findings underline that (at least in the context of peer punishment) group outcomes crucially depend on the presence of pro-social punishment types. To the best of our knowledge, our paper is the first to present causal evidence on this link. The results complement recent studies that have hinted at the importance of individuals’ inclination to punish. Ones and Putterman (2007) rank lab subjects according to a composite index, which is based on previous contribution and punishment decisions in a repeated VCM. Using the ranking to form homogenous groups of similar types, they find that subsequent cooperation is higher in groups with ‘higher-ranked’ subjects, i.e., among individuals that tend to be more cooperative and/or willing to engage in pro-social punishment.

Studying field data, Rustagi et al. (2010) find a positive correlation between natural groups’ success in managing forest commons and the number of conditional cooperators in the respective groups. They attribute this to the difference between conditional cooperators and selfish persons in their self-reported statements about time spent on forest patrols. In a similar vein, the correlational analyses by Kosfeld and Rustagi (2015) suggest that these natural groups are also better at managing forest commons if the corresponding leader’s third-party punishment behavior, as measured in a lab experiment, promotes equality and efficiency rather than being arbitrary.

Rustagi et al. (2010) and Kosfeld and Rustagi (2015) focus either on cooperation or on punishment patterns, whereas Ones and Putterman (2007) combine both patterns into a single index. By contrast, Falk et al. (2005) study both individual punishment and cooperation behavior in isolation, but without exploring the relative impact of subjects’ types on mitigating a social dilemma. They employ a strategy-method on the peer punishment-stage of a binary prisoner’s dilemma-game between three persons and relate the punishment pattern to the subject’s actual cooperation decision in the prisoner’s dilemma. While the fraction of people who cooperate and punish is similar to what we find, it differs for those who defect and punish. To some extent, this is driven by the marked amount of anti-social punishment in their data. In parts, though, this might also be due to the fact that they use the actual decision (cooperate or defect) rather than eliciting cooperation types via a strategy-method. After all, a defector might either be a selfish individual or a conditional cooperator that expects the other person to defect. Our two-dimensional type classification suggests that this distinction makes a difference for pinning down the linkage between cooperation and punishment patterns.

The authors conclude that “better forest management outcomes are not only a result of conditional cooperators being more likely to abide by the local rules of the group but also being more willing to enforce these rules at a personal cost” (p.964). The systematic causal evidence provided in this paper confirms this line of reasoning.
The classification of individuals along two dimensions offers additional insights on how the interplay of different behavioral types drives group outcomes. Accounting for the heterogeneity in punishment types significantly improves our ability to explain the large and persistent differences in cooperation across groups. Moreover, the identification of systematically different punishment patterns at the individual level provides a novel contribution to the literature which has mainly focused on variation in punishment and cooperation patterns at the aggregate level.\(^3\) Our analysis complements these studies of group-level heterogeneity and thus constitutes a potential micro-foundation that might prove useful for future studies.

It seems natural to explore several follow-up questions, especially (but by no means exclusively) in the context of decentralized sanctioning and norm enforcement. Knowledge about individuals’ (punishment) types might, for instance, help to better explain the effectiveness of other institutional arrangements aimed at sustaining cooperation (see, e.g., our work on centralized punishment in Kube and Traxler, 2011). Currently, this literature is strongly focusing on how different contribution types are affected by, or react to, the institutions at hand (e.g., Brekke et al., 2011). However, it might be worthwhile to extend this line of thinking to include punishment types as well — namely as soon as the institution at hand relies on some form of mutual monitoring, expression of preferences over certain norms, or other mechanisms that are likely to appeal differently to different kind of punishment types. Moreover, if institutions need to be adapted endogenously (e.g., via elections as in Kosfeld et al., 2009, Hamman et al., 2011, or Kube et al., 2015, or via voting by feet as in Gürerk et al., 2006), information about a population’s type composition might allow to anticipate the support for an institution for a given population. Finally, knowledge about individuals’ cooperation and punishment types might offer new solutions to optimal team composition problems (e.g., Burlando and Guala, 2005; Gächter and Thöni, 2005; Ones and Putterman, 2007).

The remainder of this paper is structured as follows. The next section discusses the design and implementation of the experiment. Section 3 presents the results from the classification of cooperation and punishment types. Section 4 shows how the presence of these different types influence group outcomes and individual behavior in a repeated game. Section 5 concludes.

2 Design and Procedures

Our experiment consists of three independent games: (1) a one-shot public-goods game without punishment (C-game), which allows us to identify individual cooperation patterns in the tradition of Fischbacher et al. (2001); (2) a one-shot public-goods game with peer punishment

\(^3\)Consider, for instance, Herrmann et al. (2008), who compare behavior in public-good games with peer punishment across 16 countries, or Henrich et al. (2006), who study third-party punishment in 15 diverse populations and observe at the aggregate level that “costly punishment positively covaries with altruistic behavior across populations” (p.1767).
(P-game) that uses a strategy-method at the punishment stage to elicit individual peer punishment patterns; and finally (3) a 10-period public-goods game with peer punishment (R-game). In the latter, random assignment produces heterogenous group compositions of cooperation and punishment types, as elicited from the C-game and P-game. We exploit this heterogeneity to analyze the interplay between the different types in the R-game and the impact on groups’ abilities to overcome social dilemmas. In addition to these three games, subjects answered a brief questionnaire.

2.1 C-Game

The C-game is a standard one-shot linear public-goods game (VCM) with the strategy-method from Fischbacher et al. (2001). Subjects are randomly assigned into groups of four. Each subject $i \in \{1, \ldots, 4\}$ is endowed with 20 tokens and decides how many tokens to contribute to the public good, $g_i$, and how many to keep for herself, $20 - g_i$. Each token allocated to the public good yields a marginal per capita return of 0.4. The payoff function is given by

$$\pi_i^C = 20 - g_i + 0.4 \sum_{j=1}^{4} g_j. \quad (1)$$

Under the assumptions of rational payoff-maximizing behavior, contributing zero is the dominant strategy of the one-shot game. In contrast, the social optimum consists of all players contributing their entire endowment to the public good.

Following the procedure of Fischbacher et al. (2001), subjects are first asked to make an unconditional contribution decision, $g_i$. Using the strategy-method, subjects then make their conditional contribution decisions. They have to indicate their contribution for all 21 possible whole numbers of average contributions among the other group members, $\bar{g}_j := \frac{1}{3} \sum_{j \neq i} g_j$, with $\bar{g}_j \in \{0, 1, \ldots, 20\}$ rounded to integers. After all decisions are made, one group member is randomly drawn. For this subject, the conditional contribution decision is implemented based on the average unconditional contributions of the other three group members. Contributions and payoffs are revealed to the subjects only at the end of the experiment.

2.2 P-Game

The P-game is a one-shot linear public-goods game with costly punishment (Fehr and Gächter, 2000, 2002). At the first stage of the game, subjects make their contribution decision, facing the same parameters as described above for the C-game. At the second stage of the P-game, each subject $i$ can assign a maximum of 10 punishment points to the other group members $j \neq i$, $0 \leq d_{ij} \leq 10$. Punishment is costly. Assigning one punishment point costs one token for the
punisher and reduces the payoff of the punished subject by three tokens (Fehr and Gächter, 2002; Herrmann et al., 2008). The payoff function is

\[ \pi_i^P = 20 - g_i + 0.4 \sum_{j=1}^{4} g_j - 1 \sum_{j \neq i} d_{ij} - 3 \sum_{j \neq i} d_{ji} . \] (2)

A fully rational, selfish agent would not engage in any punishment at the second stage of the game. Hence, contributing zero would again be the dominant strategy.

While Fehr and Gächter (2000, 2002) and the subsequent literature let subjects decide on the punishment levels for others’ actual contributions, we implement a modified strategy-method at the punishment stage.

4The strategy-method confronts subjects with a sequence of contribution triples: each subject \( i \) faces 11 screens, where each screen \( s \) presents one triple \( \{g_s^j, g_s^k, g_s^l\} \), with \( j \neq k \neq l \neq i \) and \( s \in \{1, \ldots, 11\} \). One of the 11 triples comprises the actual contributions of the other group members. The other ten triples are hypothetical combinations of contributions, each being randomly drawn from a pre-defined set of combinations (see below). All 11 triples are then presented in randomized order. For each triple, a subject has to decide how many punishment points (if any) to allocate to the other subjects.

As we aim at identifying punishment patterns at the individual level, we wanted to assure that subjects face combinations of contributions that cover different parts of the vast strategy space (up to \( 21^3 \) potential triples). To do so, we partitioned contributions into three intervals: low (L), intermediate (M), and high (H) contributions with \( g^L \in \{0, \ldots, 4\} \), \( g^M \in \{5, \ldots, 15\} \), \( g^H \in \{16, \ldots, 20\} \). We then considered the ten resulting combinations of low, intermediate and high contributions:

\[
\begin{align*}
\{g^L, g^L, g^L\} & \quad \{g^M, g^L, g^M\} & \quad \{g^L, g^M, g^H\} & \quad \{g^L, g^M, g^H\} \\
\{g^L, g^H, g^H\} & \quad \{g^M, g^M, g^H\} & \quad \{g^M, g^M, g^H\} & \quad \{g^H, g^H, g^H\}
\end{align*}
\]

Within each of the ten contribution combinations, we randomly generated eight different triples (see Appendix A1 for further details). For all 10 contribution combinations, a subject would then face one of these triples.5 Following this protocol, we observe \( 3 \times 11 \) punishment decisions for each subject.

It is common knowledge that ten out of the 11 triples are hypothetical and that only the punishment decisions for the real contribution triple become payoff relevant. However, subjects

---

4This strategy-method was first used in Kube and Traxler (2011). It can be seen as an instance of the ‘Conditional Information Lottery’ introduced by Bardsley (2000), who used it at the contribution stage of the game. For a related but different approach, see Cheung (2014) and Kamei (2014).

5One subject might see, for instance, \( \{0, 0, 0\} \) for the combination \( \{g^L, g^L, g^L\} \) and \( \{0, 2, 8\} \) for \( \{g^L, g^L, g^M\} \). A different subject might face \( \{0, 2, 3\} \) for the former and \( \{0, 2, 14\} \) for the latter. Balancing tests indicate that randomization at the individual level was successful.
neither know which one is the ‘real’ triple, nor do they know the procedure to generate the hypothetical triples. Only at the end of the experiment, the actual contribution triple and punishment choices are revealed.

2.3 R-Game

The R-game is a public-goods game with costly peer punishment (Fehr and Gächter, 2000) that is played repeatedly for ten periods. The payoff function is equivalent to the one from the P-game, summarized in equation (2). Subjects play the R-game either under a stranger (R_s) or under a partner protocol (R_p). At the beginning of the R-game, players are randomly assigned into groups of four (partner protocol, with partners not identifiable between periods) or matching-groups of eight (stranger protocol) and remain in these groups for all 10 periods. In the stranger protocol, subjects are randomly re-matched each period within their respective matching-group.

2.4 Implementation

We evaluate data for 628 subjects that participated in 29 sessions. The large sample allows us to study the role of heterogenous group compositions for group outcomes (see Section 4). For each subject we observe 21 conditional contribution decisions in the C-game, 3 \times 11 punishment decisions in the P-game as well as 10 contribution and 30 punishment decisions in the R-game. 452 subjects played the R-game under a partner protocol, 176 subjects under a stranger protocol. The experiments were conducted at the University of Bonn’s BonnEconLab, using the experimental software zTree (Fischbacher, 2007). Subjects were recruited online using Orsee (Greiner, 2015). To prevent strategic spillovers between games, subjects received the instructions for each subsequent part of the experiment once the previous part had been completed by all participants. Standard experimental procedures were followed. Results and payoffs from the C- and the P-game were only revealed at the end of the experiment. Results and payoffs from the R-game were revealed after each period. Including a follow-up questionnaire, a session lasted approximately 100 minutes. On average, subjects earned 19.88 Euro, including a 5 Euro show-up fee.

6Testing whether subjects punish the (unknown) real versus the hypothetical contributions differently, we find no significant differences whatsoever.

7The instructions under the stranger protocol explicitly informed subjects that groups are randomly re-shuffled each period, without stating the total number of matching groups per lab session. (One session typically consisted of 24 subjects spread over three matching-groups.) The implementation of multiple matching groups per session is a common practice that balances the benefits from reducing reputational concerns and total implementation costs. Moreover, since parts of our analyses exploit between matching-group variation in types, smaller matching-groups are important to obtain sufficient between group variation.

8The instructions and further details on the procedure are available in the Online Appendix.
3 Individual Patterns in Cooperation and Punishment

This section studies individual punishment (Subsection 3.2) and cooperation patterns (3.3). Section 3.4 analyzes if and how these two patterns are aligned. Before doing so, we first discuss behavioral predictions based on the related literature.

3.1 Behavioral Predictions

As the VCM with punishment is well studied in the literature, the baseline predictions (as well as their limited predictive power) are well-known: As already noted above, the subgame-perfect Nash equilibrium for selfish payoff-maximizing players is that nobody contributes to the public good at the first stage since players anticipate that nobody will engage in costly punishment at the second stage. In that case, we should observe contribution and punishment profiles that are both ‘flat’ at zero, i.e., free-riders that do not punish.

Concerning the contribution profiles, previous studies have found a significant number of conditional cooperators. The seminal paper by Fischbacher et al. (2001), for instance, classifies 50% of subjects in their sample as conditional cooperators and 30% as free-riders. Similarly, heterogeneity has also been observed with respect to punishment behavior. For example, the between-group comparison in Herrmann et al. (2008) reveals significant differences in both the level and the targeting of punishment. Correspondingly, heterogeneity in individual punishment behavior is observed in Falk et al. (2005), Kube and Traxler (2011), Cheung (2014), and Kosfeld and Rustagi (2015). Although these articles differ in their specific elicitation methods and underlying games, they all point to the existence of different punishment types: some subjects do not use punishment, while others do engage in costly punishment. Among the latter, some apply punishment in a pro-social manner (targeting free-riders or low cooperation levels) others punish anti-socially (e.g., targeting high cooperation levels).

In light of the previous evidence, we therefore expect to observe individual-level differences in contribution and punishment patterns, too. However, it still remains open if and how individual contribution and punishment patterns are related.

Conceptually, one might argue that these patterns should be closely aligned, since both peer punishment and voluntary contributions (in the C-game) constitute a cooperation problem. One should thus expect that free-riders do not punish (as long as no monetary gains are expected to arise from punishment, an argument already made by Oliver, 1980). Similarly, Fehr and Gächter (2000, p.984) postulate that conditional cooperators are willing to engage in the costly punishment of free-riders, arguing that this would be predicted by models of reciprocity and equity. In fact, Leibbrandt and López-Pérez (2012, p.762) find that a “combination of inequity-averse and selfish types can sufficiently capture ... punishment patterns”, as measured by conditional responses to a dictator’s choices in ten binary allocation decisions. Likewise, the results in Che-
ung (2014, p.130) on the determinants of peer punishment “are directionally consistent with the predictions of the Fehr and Schmidt (1999) model of inequality aversion.”

Similar notions of a close alignment of cooperation and punishment are developed in many other contributions. The review by Gächter and Herrmann (2009) on human cooperation, for instance, refers to voluntary contributions as positive and to punishment as negative reciprocity. Likewise, Ones and Putterman (2007, p.498) implicitly assume that positive and negative reciprocity are “two sides of the same coin” when they explore the impact of punishment and contribution behavior on group outcomes. A very different view is offered by Peysakhovich et al. (2014), who use factor analyses to compare individual decisions across six different one-shot games. They conclude that “punishment and cooperation may be separate phenomena, rather than being driven by a common altruistic motivation” (p.2) and thus “may not be two sides of the same coin” (p.3). If their findings were to extend to cooperation and punishment behavior within a given situation, it might be that conditional cooperators do not necessarily engage in (pro-social) punishment. Likewise, it might be that free-riders do spend resources on (pro-social) punishment.

To wrap-up: on the basis of the existing evidence we clearly expect to observe heterogeneity in both contribution and punishment patterns. However, the literature offers competing conjectures regarding the outcome of the two-dimensional classification approach: behavioral patterns might be fully aligned or (at least partially) diverging. Our two-dimensional classification offers an explorative approach that seeks to clarify whether or not cooperation and punishment are indeed two sides of the same coin. While we can offer a novel take on this question, we do not offer a specific test of the underlying motivation. To convincingly sort out different channels emphasized in competing models (e.g., of other-regarding preferences), one would require much richer data on individuals’ beliefs and decisions under very different payoff functions (i.e., from different games and different parametrization).

3.2 Individual Peer-Punishment Patterns

3.2.1 Primary Classification of Punishment Types

In a first attempt to classify individual peer-punishment patterns, we model punishment $d_{ij}$ as a linear function of player $j$’s contribution to the public good (with $j \neq i$):

$$d_{ij} = \alpha_i + \beta_i(20 - g_j) + \varepsilon_i.$$  

Footnote 9: For the underlying intuition see footnote 22 below and, for a more formal analysis, the Supplementary Online Appendix of Leibbrandt and López-Pérez (2012, p.A42).

Footnote 10: Among others, the authors observe an unconditional contribution decision in a VCM and a punishment decision in a binary prisoners’ dilemma.
The regressor in eq. (3), 20 − gj, is j’s deviation from contributing the full endowment (20 tokens). This linear transformation will facilitate the interpretation of the coefficients (see below).\textsuperscript{11} Using the data from the strategy-method in the P-game (for the punishment decisions of the one-shot game), we separately estimate α_i and β_i for each of our 628 subjects. The estimated coefficients capture individual-level heterogeneity in punishment patterns.

It is important to realize that conventional observational data would not allow for a proper identification of the coefficient β_i at the individual level. In one-shot public good games with peer punishment, one would only observe three punishment choices per subject. Similarly, in repeated games like our R-game, contributions shape punishment and punishment shapes contributions simultaneously.\textsuperscript{12} Our strategy-method breaks this simultaneity by introducing exogenous variation in gj. Following this line of reasoning, we focus on the subjects’ punishment choices for the 10 × 3 exogenous contribution triples of the P-game, i.e., we exclude the triple with the actual contributions, leaving us with 30 observations per subject.\textsuperscript{13}

Running 628 regressions with \( N_i = 30 \), we obtain the estimates \( \hat{\alpha}_i \) and \( \hat{\beta}_i \) (along with robust standard errors) for each subject. Based on these estimates, we then classify the subjects’ punishment patterns. Our classification distinguishes between subjects that do not punish, ‘pro-social’, and ‘anti-social’ punishers:

1. A subject is classified as a ‘Non-Punisher’ (NPun) if she assigns zero punishment points in each case, i.e., \( d_{ij} = 0 \) for all \( g_j \). In equation (3), this is depicted by \( \hat{\alpha}_i = \hat{\beta}_i = 0 \).

2. Subjects that target their punishment towards those that contribute little or nothing to the public good have a punishment pattern that is upward sloping in \( (20 - g_j) \). These subjects, with \( \hat{\beta}_i > 0 \) and \( p \leq 0.01 \), are classified as ‘Pro-social Punishers’ (Pun).

3. Subjects are classified as ‘Anti-social Punishers’ (APun), if their punishment is either increasing in the other’s contribution \( g_j \), i.e., if \( \hat{\beta}_i < 0 \) and \( p \leq 0.01 \), or if they display a significant positive but unsystematic level of punishment: \( \hat{\alpha}_i > 0 \) with \( p \leq 0.01 \) and an insignificant slope coefficient \( \hat{\beta}_i \) with \( p > 0.01 \).\textsuperscript{14}

\textsuperscript{11} Estimating a model with \( d_{ij} = \alpha'_i + \beta'_i g_j + \varepsilon'_i \) would yield equivalent estimates with \( \hat{\beta}_i = -\hat{\beta}'_i \).

\textsuperscript{12} Due to serial correlation in choices within subjects and (matching-)groups, one cannot easily avoid endogeneity problems (e.g., by using lagged values). In fact, our classification approach produces quite different results if we use the exogenous variation from our strategy-method or the endogenous variation in the repeated game data (see Table S.5 in the Online Appendix).

\textsuperscript{13} Our results are insensitive to including the three punishment decisions for the real contribution triple.

\textsuperscript{14} The literature typically defines anti-social punishment in reference to a subject’s own contribution, i.e., if the punishment-receiving subject contributed a larger or equal amount to the public good compared to the punishing individual (e.g., Herrmann et al., 2008). Our primary classification approach deviates from this self-centered notion of anti-social punishment as we do not consider punisher i’s own contribution \( g_i \). Still, APun-types reflect patterns of punishment that are targeted towards high contributors.
Punishment patterns that cannot be assigned to one of these three types are summarized in a group of non-classified (NCL) patterns. The different types and their stylized punishment patterns are illustrated in Figure A1 in the Appendix.

The results from our classification approach are presented in Figure 1. 47.1% of our subjects are classified as pro-social punishers, 40.3% are non-punishers, 2.6% display an anti-social pattern, and 10.0% are in the residual group of non-classified patterns (NCL). Subjects from the latter group show very low levels of sporadic punishment (as illustrated in Figure A1). In fact, if we relax the strict definition of NPun to include also subjects with \( \hat{\alpha}_i \approx \hat{\beta}_i \approx 0 \), then every single NCL type would be re-classified as NPun. These (de-facto) non-punishers would then account for 50.3% of the sample.\(^{15}\)

Figure 1: Primary Punishment Types and Patterns

<table>
<thead>
<tr>
<th>Type</th>
<th>N</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pun</td>
<td>296</td>
<td>47.1</td>
</tr>
<tr>
<td>NPun</td>
<td>253</td>
<td>40.3</td>
</tr>
<tr>
<td>APun</td>
<td>16</td>
<td>2.6</td>
</tr>
<tr>
<td>NCL</td>
<td>63</td>
<td>10.0</td>
</tr>
<tr>
<td>Total</td>
<td>628</td>
<td>100.0</td>
</tr>
</tbody>
</table>

Notes: Punishment type distribution and average punishment patterns (in the 20 \(- g_j\)-space) for the different types: pro-social punishers (Pun), non-punishers (NPun), anti-social punishers (APun), and non-classified punishment profiles (NCL). To ease illustration, the pattern for the latter is not plotted.

The results show that our sample is characterized by a high frequency of Pun types. The average punishment pattern, indicated by the dashed black line in Figure 1, is therefore clearly increasing in 20 \(- g_j\). Note further that the slope of the punishment pattern is relatively steep. The average \( \hat{\beta}_j \) among Pun types is 0.135 (the median is 0.124). This suggests that a player \( j \) — who faces an average Pun type — receives around 0.14 punishment points for a one unit decline in her contribution \( g_j \). If player \( j \) faces two [or even three] Pun types in her group, the marginal punishment increases to 0.28 [0.42] points. Given the parameters of the game (see equation 2) this translates into marginal costs of 0.84 [1.26] token — which weakly [strongly] dominates the marginal payoff gains from free-riding (0.6 token).

---

\(^{15}\)These results are documented in the Online Appendix (see Figure S.1). It is further worth noting that we obtain very similar type distributions if we use Spearman’s rank correlation to classify punishment patterns (see Tables S.1 and S.2).
3.2.2 Robustness and Classification of Self-Centered Punishment Types

How robust are our type classifications? Note first that the strategy-method introduces, by design, random variation in $g_j$. This renders the estimated coefficients from (3) fairly insensitive to adding further control variables (e.g., controls for contributions $g_k$ and $g_l$, $k \neq l \neq j$).\(^\text{16}\) Obviously, this statement does not imply that eq. (3) is the ‘best model’ to describe individual punishment patterns. Moreover, it does not imply that we consider efficiency motives (to punish any deviation from the socially optimal behavior, i.e., from contributing the full endowment) as the primary driver of peer punishment. The point is simply that our primary classification approach yields quite robust results (see Section II in the Online Appendix).

Motivated by earlier findings in the literature, we nevertheless explore one alternative, more refined approach to classify patterns of peer-punishment. More specifically, we account for the fact that individual $i$’s disposition to punish might be ‘self-centered’ around her own contribution (from the first stage of the P-game). We thus consider a model that allows punishment patterns to differ between the domain $g_j < g_i$ and $g_j \geq g_i$:

$$d_{ij} = \alpha_i' + \beta_{i1}' \max(g_j - g_i; 0) + \beta_{i2}' \max(g_i - g_j; 0) + \epsilon_i'. \quad (4)$$

Based on the estimates from this model (in particular, $\hat{\beta}_{i1}'$ and $\hat{\beta}_{i2}'$) one can differentiate between numerous, refined patterns of punishment (see Figure A2 in the Appendix). Our refined classification will again focus on a limited set of types. (1) Subjects with a significant pro-social punishment slope in the domain $g_j < g_i$ (i.e., $\hat{\beta}_{i2}' > 0$) but an insignificant slope coefficient for $g_j \geq g_i$ (i.e., $\hat{\beta}_{i1}'$ with $p > 0.01$) are classified as ‘Self-centered Punishers’ (SPun'-types). (2) All further subjects with statistically significant, positive slope coefficients (i.e., with two positive $\beta$-estimates or only $\hat{\beta}_{i1}' > 0$) are subsumed in a class of (further) ‘Pro-social Punishers’ (Pun'-types). In addition, we distinguish between (3) ‘Non-Punishing’ (NPun'; defined as above) and (4) ‘Anti-socially Punishing’ (APun') types. The residual category are again non-classified patterns (NCL'-types).\(^\text{17}\)

Table 1 compares the results from this more refined type classification approach to our primary classification outcomes. Among the 296 subjects labeled as pro-socially punishing (Pun) according to our primary classification, 270 (more than 91%) do in-fact display a self-centered pattern of punishment (SPun'). Together with 11 further subjects (that were NCL according to our primary approach) we obtain at a total of 281 self-centered punishment patterns (44.7% of all 628 subjects). Beyond this large group, there are only six pro-social Pun'-types with patterns

\(^{16}\)A classification that builds, for instance, on the estimates $\hat{\alpha}_i$ and $\hat{\beta}_{i1}$ from the equation $d_{ij} = \alpha_i + \beta_{i1}(20 - g_j) + \beta_{i2}(20 - g_k) + \beta_{i3}(20 - g_l) + \epsilon_i$ differs for a mere 11 subjects (1.8% of our sample). Adding dummies that capture the sequence at which subject $i$ faced a certain triple does not change this picture.

\(^{17}\)In terms of the stylized illustration in Figure A2, the pattern from panel b would classify as SPun'-type; the pattern illustrated in panels a, c, and d would be all subsumed as Pun'-types; and, finally, the patterns from panels a.i to d.i would be classified as APun'-types.
that are not self-centered (as defined above). Hence, almost all pro-social punishment in our sample is self-centered and these self-centered patterns constitute the most frequent punishment type. We will return to these findings below.

Table 1: Primary and Refined (Self-Centered) Classification of Punishment Types

<table>
<thead>
<tr>
<th>Primary Classification ↓</th>
<th>Refined Type Classification</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pun'</td>
<td>SPun'</td>
</tr>
<tr>
<td>Pun</td>
<td>6</td>
<td>270</td>
</tr>
<tr>
<td>NPun</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>APun</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>NCL</td>
<td>0</td>
<td>11</td>
</tr>
<tr>
<td>Total</td>
<td>6</td>
<td>281</td>
</tr>
</tbody>
</table>

Notes: Row values display the outcome from our primary classification approach based on eq. (3). Column values depict classification results for the refined ('self-centered') approach that builds on equation (4).

Regarding the other types we observe only minor changes in the number of anti-social punishers and a modest increase in the number of non-classified patterns (from 63 in our primary to 83 in our refined classification approach). The latter observation is related to a conceptual limitation of the refined classification method: for very low or very high contributions \( g_i \), there will be few observations with \( g_j < g_i \) or \( g_j \geq g_i \), respectively. This certainly reduces the scope to obtain precise estimates for both \( \beta \) coefficients of eq. (4).

The methodological complications associated with the refined classification approach, together with the fact that pro-social punishment follows almost unanimously a self-centered pattern (which, in turn, nullifies the scope for distinguishing among different types of pro-social punishment), motivates us to work with the primary type classification throughout the remainder of the paper. Given the results from Table 1, however, it is unsurprising that the results from the subsequent analyses are qualitatively insensitive to using the type distribution from the self-centered classification approach.

3.3 Individual Cooperation Patterns

Next we analyze the strategy-method data from the C-game, where each subject states — conditional on all potential values for the others’ average contribution — how much to contribute to

\[ 18 \text{In an attempt to cope with this limitation, we applied the classification only to individuals with } 5 \leq g_i \leq 15. \] For this range of contributions, we obtain a very similar pattern as the one displayed in Table 1. A further issue concerns the unbiasedness of the estimates: including \( g_i \) in the estimation model mechanically re-introduces endogeneity. To see this point, note that all unobserved individual factors \( \psi_i \) that shape punishment \( d_{ij} \) are absorbed in the error term \( \varepsilon_i' \). As these unobserved factors \( \psi_i \) will also influence \( g_i \), we obtain \( \text{Cov}(g_i, \varepsilon_i') \neq 0 \). One can show, however, that this will mainly bias the estimates for \( \alpha_i' \).
Figure 2: Cooperation Patterns and Contribution Types

![Cooperation Patterns and Contribution Types](image)

<table>
<thead>
<tr>
<th>Type</th>
<th>N</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>CC</td>
<td>382</td>
<td>60.8</td>
</tr>
<tr>
<td>FR</td>
<td>130</td>
<td>20.7</td>
</tr>
<tr>
<td>TC</td>
<td>54</td>
<td>8.6</td>
</tr>
<tr>
<td>NC</td>
<td>62</td>
<td>9.9</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>628</td>
<td>100.00</td>
</tr>
</tbody>
</table>

Notes: The figure presents the distribution of contribution types, following Fischbacher et al. (2001) and Fischbacher and Gächter (2010), and the average cooperation patterns for the different types: Conditional Cooperators (CC), Free-Riders (FR), Triangular Contributors (TC), and Non-classified (NC) cooperation patterns. To ease illustration, the pattern for the latter is not plotted.

the public good (Fischbacher et al., 2001). Based on these data we classify individual cooperation types.

Consistent with our approach from above we separately estimate for each subject $i$ the linear model $g_i = a_i + b_i \bar{g}_j + e_i$ (with $\bar{g}_j := \frac{1}{3} \sum_{j \neq i} g_j$). Applying the type classification proposed by Fischbacher and Gächter (2010) we distinguish between Conditional Cooperators (CC, with $\hat{b}_i > 0$ at $p \leq 0.01$), Free-Riders (FR, with $g_i = 0$ for all $\bar{g}_j$, i.e., $\hat{a}_i = \hat{b}_i = 0$), Triangular Contributors (TC), and Non-classified (NC) cooperation patterns. Figure 2 presents the distribution of these types among the 628 subjects from our sample. The observed type distribution, as well as the cooperation patterns, are remarkably similar to those reported in Fischbacher et al. (2001) and Fischbacher and Gächter (2010): 61% are conditional cooperators and 21% are free-riders. The remaining 18% display a triangular or a non-systematic contribution pattern.  

3.4 Two-Dimensional Type Distribution

Finally, we combine the results from subsections 3.2 and 3.3 to arrive at a two-dimensional type classification, which links punishment and cooperation patterns at the individual level. In this vein, we can examine the relationship between individuals’ disposition to cooperate and their inclination to engage in punishment. Table 2 presents the results from the two-way classification.

---

19Classifications based on Spearman’s rank correlation (as in Fischbacher et al., 2001) yield almost identical results. See the Online Appendix for details (Table S.3).
Table 2: Two-way Distribution: Contribution and Punishment Types

<table>
<thead>
<tr>
<th>Contrib. Types</th>
<th>Punishment Types</th>
<th>Sum (%)</th>
<th>(N)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pun</td>
<td>NPun</td>
<td>APun</td>
</tr>
<tr>
<td>CC</td>
<td>33.4</td>
<td>21.2</td>
<td>1.4</td>
</tr>
<tr>
<td>FR</td>
<td>6.5</td>
<td>11.6</td>
<td>0.3</td>
</tr>
<tr>
<td>TC</td>
<td>4.3</td>
<td>2.9</td>
<td>0.0</td>
</tr>
<tr>
<td>NC</td>
<td>2.9</td>
<td>4.6</td>
<td>0.8</td>
</tr>
<tr>
<td>Sum (%)</td>
<td>47.1</td>
<td>40.3</td>
<td>2.6</td>
</tr>
<tr>
<td>Sum (N)</td>
<td>296</td>
<td>253</td>
<td>16</td>
</tr>
</tbody>
</table>

Notes: Within subject two-dimensional contribution and punishment type distribution in percent, for 628 subjects respectively. N shows the absolute type distribution per game.

The table reveals that, overall, a third of our sample (33.4%) are conditional cooperators with a pro-social peer punishment pattern (CC×Pun). Almost 12% are free-riders in the C-game that do not punish in the P-game (FR×NPun). In addition to these types with aligned patterns, we also observe a non-trivial fraction of subjects with diverging patterns: 21% of all subjects are conditional cooperators that do not punish at all (CC×NPun) and more than 6% are free-riders with a pro-social punishment pattern (FR×Pun).

A different way of presenting the distribution of these four types — which cover almost three out of four subjects in our sample — is provided in Figure 3. The bar graphs indicate that roughly every second conditional cooperator punishes pro-socially (55%) and that more than one out of two free-riders do not punish at all (56%). In addition to these types, whose cooperation and punishment patterns are aligned, there seems to be a second archetype of subjects with diverging patterns: every third (35%) conditional cooperator does not punish and, analogously, almost one in three (32%) free-riders punishes pro-socially. Hence, the overlap between conditionally cooperative and (pro-social) punishing individuals is far from perfect.

3.4.1 Further Analyses

In a next step, we examine the distribution of the underlying coefficients of the type classifications (in particular, \( \hat{\beta}_i \) and \( \hat{b}_i \); see Figure A3 in the Appendix). The analysis reveals a positive correlation between \( \hat{b}_i \) and \( \hat{\beta}_i \): ‘stronger’ conditional cooperators tend to have ‘steeper’ punishment patterns. However, the correlation is again far from perfect. Among CC×Pun-types, for instance, we observe an insignificant correlation coefficient of \( \rho = 0.094 \) (\( p = 0.173 \)).

20The Spearman correlation is slightly stronger (0.128) and statistically significant (\( p = 0.064 \)).
Notes: The graph depicts the conditional frequency of Pun- and NPun-types among conditional cooperators (CC) and free-riders (FR), respectively.

Making use of the data from our questionnaire, we further studied whether individual characteristics, personality traits (big five, etc.) and attitudes (risk, trust, etc.) correlate with the contribution and punishment types (extensive margin variation) or patterns within types (intensive margin variation). Our analysis reveals three strong and robust predictors for the type assignments. First, we find that subjects who express their willingness to impose social sanctions on norm violators among their peers (e.g., drunk drivers; see the survey questions in Traxler and Winter, 2012) are significantly more likely to be pro-social punishers (Pun-types). This observation suggests that the survey measure on norm enforcement is consistent with the behavioral measure that builds on the observed pattern of peer punishment. Second, we find that subjects who see themselves as more reserved (see Rammstedt and John, 2007), are much more likely to be a NPun-type. Third, considering the different contribution types, we detect a strong gender effect: females have a much higher likelihood of being a conditional cooperators and, vice versa, a much lower probability of being a free-rider.21

To study intensive margin variation within types, we examined correlations of observables with the slopes of the subjects’ contribution and punishment patterns ($\hat{\beta}_i$ and $\hat{b}_i$). Our analysis reveals that, among Pun-types, the slope of the punishment pattern is lower for females as well as for subjects with a high level of agreeableness in the big five (Rammstedt and John, 2007). For the cooperation patterns of CC-types, we find that those who express a high level of trust in others have a steeper contribution pattern: they are more likely to one-to-one match others’ contributions.

21Probit and LPM estimates underlying these results are available from the authors.
3.4.2 Interim Summary

Summing up, our two-dimensional classification reveals the existence of two behavioral archetypes. First, the largest group of subjects is characterized by an overlap in their pro-social behavioral patterns. This group includes conditionally cooperative types that do engage in pro-social punishment ($CC \times Pun$) and free-riders that do not invest in punishment at all ($FR \times NPun$). For these types, cooperation and punishment are indeed “two sides of the same coin” (Ones and Putterman, 2007).

Second, our analysis also identifies a significant share of individuals that are conditional cooperators which do not punish ($CC \times NPun$) as well as free-riders that are classified as pro-social punishers ($FR \times Pun$). About every third conditional cooperator and, analogously, every third free-rider displays such a divergence in cooperation and punishment patterns. The identification of this second archetype therefore suggests that cooperation and punishment may indeed be separate phenomena — at least for some individuals (see Peysakhovich et al., 2014). While the latter finding seems interesting in itself, it further implies that individual inclinations to cooperate and to punish are far from perfectly correlated in our sample. We can thus assess the interplay between the different types and their role for explaining outcomes in another independent situation: the R-game.

4 Group Composition and Contributions in the Repeated Game

In this section we demonstrate the benefits from identifying heterogenous punishment types for explaining group and individual level heterogeneity in repeated public goods games with peer punishment. To this end, we exploit the data from the 10 periods of the $R_p$- and the $R_s$-game (partner and stranger design, respectively). We analyze the influence of group compositions on group outcomes and individual behavior. Motivated by other studies which document the benefits from grouping pro-social individuals in a repeated VCM (e.g., Gächter and Thöni, 2005; Ones and Putterman, 2007), we start out by computing the number of conditional cooperators ($CC$) and pro-social punishers ($Pun$) for each group (and for each matching group of eight in $^2$One aspect that is beyond the scope of the present paper is the explanation of this second archetype based on existing theories of other-regarding preferences. Self-evident models to structure our data are based on theories of inequality aversion, in particular Fehr and Schmidt (1999) (F/S). (Obviously, we do not estimate coefficients from self-centered models of punishment. As pointed out in Section 3.2.2, the overall picture from our type classifications hardly changes for these more complex models.) Intuitively speaking, in F/S the decision to contribute is shaped by the aversion against advantageous inequality (i.e., the parameter $\beta$ in F/S), whereas pro-social punishment is motivated by aversion against disadvantageous inequality (i.e., the parameter $\alpha$). As such, F/S can easily accommodate the ‘aligned’ type combinations $CC \times Pun$ (high $\alpha$ and $\beta$) and $FR \times NPun$ (low $\alpha$ and $\beta$). Given the specific parameters of our experiment (4 players, MPCR of 0.4, and punishment technology of 1:3), also the less intuitive $CC \times NPun$-type is consistent with F/S-subjects with a sufficiently strong aversion against advantageous inequality but only a mild aversion against disadvantageous inequality. Yet, using F/S to explain the combination of free-riders that punish others with low-contributions ($FR \times Pun$) is not that straightforward and would require assumption regarding (players’ expectations about) the distribution of the parameters $\alpha$ and $\beta$ in the population.}
the stranger design). Making use of the random assignment of subjects into groups — a point which is discussed in detail in the Appendix A2 (see also Table A1) — we first evaluate the impact of having more or less CC- or Pun-types on a group’s average contribution level.

Figure 4: Average (Matching)-Group Contributions by Type Prevalence

Partner (R_p-game)          Stranger (R_s-game)

(A) Variation in CC-Types   (B)

(C) Variation in Pun-Types  (D)

Notes: Panels A and B show the average contribution per period among the (matching)-groups for different frequencies of CC-types within the respective group. Similarly, panels C and D depict differences in average contributions for different numbers of Pun-types. Panels A and C consider the partner design (R_p, with groups of four subjects), panel B and D are based on the stranger design (R_s, with eight-player matching groups). The underlying variation of types across (matching-)groups is presented in Table A1 in the Appendix.

4.1 Descriptive Evidence

A first glimpse at the results is provided by Figure 4. It depicts the average contribution per group over 10 rounds for different group compositions.\textsuperscript{23} Panel A [B] compares contributions for

\textsuperscript{23}To ease exposition, the figure pools groups with similar type compositions. The raw data are illustrated in the Online Appendix (see Figure S.4).
[matching-] groups with different numbers of $CC$-types. The figure shows a strong positive relationship between the number of $CC$-types and the average contribution level — an observation that is fully in line with the results from Gächter and Thöni (2005).

Panel C [and D] compares [matching-] groups with different numbers of $Pun$-types. Similar as above, we observe that contributions are higher in groups that contain more pro-social punishers. However, the standard errors are now smaller and, what is more important, average contribution in ‘good’ groups are higher in panel C as compared to panel A: During the last 5 periods of the $R_p$-game, groups with 3 or 4 $Pun$-types have an average contribution of 17.2 tokens. Groups with 3 or 4 $CC$-types ‘only’ reach 14.9 tokens on average. The difference is significant at the 5%-level ($p = 0.036$ in a two-sided t-test).

In the stranger design, we generally observe lower contribution levels. Comparing panel B and D further shows that the differences among ‘top’ groups are less pronounced than in the $R_p$-game. Matching-groups with either few $CC$- or few $Pun$-types show strongly declining contributions over time, a pattern well documented for repeated public goods games without punishment.\footnote{Figure S.2 in the Online Appendix replicates Figure 4 for average group payoffs rather than contributions. This exercise delivers similar findings as those discussed above.}

4.2 Regression Analysis: Group Contributions

Figure 4 shows that the number of both $CC$- and $Pun$-types are important determinants of average contributions at the group level. To investigate the role of the different types in more detail, we conduct a regression analysis. We estimate models of the structure

$$\bar{g}_{lt} = \gamma_0 + \gamma_1 CC_{\text{few}} + \gamma_2 CC_{\text{many}} + \sum_t \delta_t D_t + \epsilon_{lt},$$

(5)

where $\bar{g}_{lt} := \frac{1}{n} \sum_{i=1}^{n} g_{il}$ is the average contribution in group $\ell$ in period $t$. The explanatory variables are dummies indicating if there are few (one or two) or many (three or four) $CC$-types in a group.\footnote{The reference category are groups with zero $CC$-types. In the interpretation of the point estimates discussed below, one should keep in mind that most groups are populated by at least some conditional cooperators (see Table A1 in the Appendix).} In addition, the specification accounts for period-fixed effects. The results from linear random-effects estimations of equation (5) for the 113 groups in the partner design ($R_p$-game) are presented in column (1) of Table 3.\footnote{Tobit estimations yield almost identical results (see the Online Appendix, Table S.7).}

Consistent with the graphical evidence from above, and again in line with Gächter and Thöni (2005), the estimates document that groups with a higher number of conditional cooperators achieve higher contributions. The point estimates indicate that groups with one or two $CC$-types reach contributions which are, on average, around 4 tokens higher than in groups with zero $CC$
types. For groups with three or four CC types, this difference increases to 7 tokens. In economic terms, both coefficients are sizeable. Statistically speaking, however, the first coefficient, which corresponds to $\gamma_1$ from equation (5), is only weakly significant. A Wald test further rejects $\gamma_1 = \gamma_2$ with $p = 0.003$.

Column (2) reports the results for a model that uses dummies indicating groups with few or many Pun- (rather than CC-)types. The point estimates are of similar magnitude but the coefficients are more precisely estimated: on average, a group with one or two [three or four] Pun-types achieves contribution levels that are around 4 [7] tokens above those observed for groups with zero Pun-types. Both dummies are now significant at the 1% and 5% level, respectively (with the two estimates being significantly different from each other; $p = 0.000$). Note further that all information criteria reported in Table 3 indicate that the estimated model in column (2) clearly dominates the one from column (1): the $R^2$ strongly increases and the Akaike information criterion (AIC) declines, indicating a better model fit. This underlines the usefulness of information about the number of Pun-types in a group for explaining the heterogeneity in cooperation levels between groups.

The last point is further corroborated by the outcome reported in column (3). The specification includes both sets of dummies from before and thus directly assesses the relative importance of having more or less CC- or Pun-types in a group. These two dummies are certainly correlated; nevertheless, the significant share of subjects with diverging cooperation and punishment patterns (see above) in combination with our fairly large sample allows us to distinguish the role of CC- and Pun-types.

The results reported in column (3) show that the model clearly yields a better fit than the one using only information about CC-types (column 1); however, $R^2$ and AIC only improves modestly as compared to the specification from column (2). Put differently: once we account for a group’s Pun-types, adding information about CC-types only weakly increases explanatory power. The results further indicate that the estimated coefficients on the two CC-dummies shrink in magnitude while standard errors increase: one coefficient ($\gamma_1$) loses statistical significance, the other one ($\gamma_2$) remains significant at the 5% level. The precision of the two Pun-dummies decreases slightly, too; however, both coefficients remain significant at the 1% and 10% level, respectively.

The last specification, presented in column (4), adds dummies for the prevalence of CC×Pun-types (in the spirit of an interaction term). The outcome shows that, for a given number of CC- and Pun-types, having more or less of these two-way types does not matter for the groups’ average contribution levels. In fact, the AIC suggest that the simpler specification from column (3) dominates the one from (4). Concerning the other type dummies, it is reassuring
to see that the estimates are almost unchanged — an observation that is consistent with the random assignment of subjects to groups.27

In a next step, we consider the data from the stranger design. Columns (5)–(8) in Table 3 present the estimation output from an analogous set of regressions as those discussed above. The results are similar to those for the partner design. Again, we observe that a higher number of CC- or Pun-types within a matching group is associated with higher average contributions. Similar as above, specification (6), which controls for variation in the number of Pun-types, has a higher explanatory power and a better fit than specification (5). In column (7), when we add dummies for both types, only the ones on the Pun-types remain significant. In addition, the point estimates for the CC-dummies become much smaller now. In fact, post-estimation tests for specifications (7) and (8) both reject $CC_{\text{few}} = \text{Pun}_{\text{few}}$ ($p = 0.026$ and $p = 0.069$, respectively)

27Post-estimation tests following specifications (3) and (4) reject $\gamma_1 = \gamma_2$ ($p = 0.013$ and $p = 0.039$, respectively) and, analogously, the equality of the two Pun-dummies ($p = 0.000$). However, we cannot reject $CC_{\text{few}} = \text{Pun}_{\text{few}}$ and $CC_{\text{many}} = \text{Pun}_{\text{many}}$ ($p = 0.732$ and $p = 0.708$, respectively).
as well as $CC^{\text{many}} = Pun^{\text{many}}$ ($p = 0.026$ and $p = 0.089$, respectively). The analysis therefore confirms the picture from above: in the presence of peer punishment, having more $Pun$-types in a group seems to be key for achieving high contribution levels, in particular, in the stranger design.

### 4.3 Regression Analysis: Individual Contributions

Above we showed how variation in groups’ type composition affects average group contributions. We now turn to the underlying individual behavior that is driving these results. To investigate the influence of the group composition on individual contribution decisions, we estimate the equation

$$g_{it} = \lambda_0 + \lambda_1 CC_{i}^{\text{few}} + \lambda_2 CC_{i}^{\text{many}} + \lambda_3 Pun_{i}^{\text{few}} + \lambda_4 Pun_{i}^{\text{many}} + \phi Pun_i + \sum_t \delta_t D_t + \epsilon_{it}, \quad (6)$$

The first set of dummies now captures whether individual $i$ faces few or many $CC$- or $Pun$-types among the other players in her group $t$. The $\lambda$-coefficients thus reflect the impact from variation in the type composition among $i$’s peers on her contribution. The model further includes a dummy $Pun_i$, which indicates if $i$ has been classified as a $Pun$-type herself. As an alternative, we will consider the dummy $NPun_i$, which indicates that she did not punish in the P-game. The coefficient $\phi$ then captures whether being a $Pun$ (or $NPun$) type is correlated with higher or lower contributions. Finally, note that we estimate equation (6) separately for subjects classified as free-riders ($FR$) and conditional cooperators ($CC$). Considering these two groups separately allows for type-specific responses to variation in the group composition. Moreover, any unconditional differences among these two contribution types will be reflected in different constants ($\lambda_0$).

The results from estimating eq. (6) for the partner design are presented in Table 4. Let us first focus on the estimates for conditional cooperators. Columns (1) and (2), which present specifications that separately include either the $CC_{i}$ or the $Pun_{i}$ dummies, suggest that a $CC$-type’s contribution increases with the number of (other) conditional cooperators as well as with the number of pro-social punishers in the group: post-estimation tests reject $\lambda_1 = \lambda_2$ ($p = 0.080$) and $\lambda_3 = \lambda_4$ ($p = 0.001$). In terms of statistical and economic significance, however, an increasing number of $Pun$-types seems to exert a much stronger effect on contributions. This point is also documented in column (3), where the $CC_{i}$ dummies become statistically insignificant, whereas

More precisely, in the partners protocol, the dummies capture if there are few (one) or many (two or three) $CC$- or $Pun$-types among the other three players in the group. For the strangers design, the dummies with superscript few [many] indicate that two to four [five or more] subjects out of the seven other players in the matching group were classified as $CC$- or $Pun$-type, respectively.
<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dependent variable:</strong> Individual Contribution ((g_{it}))</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Conditional Cooperators (CC)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CC\textsuperscript{few}</td>
<td>1.464</td>
<td>0.551</td>
<td>1.673</td>
<td>1.925</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(1.544)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CC\textsuperscript{many}</td>
<td>3.089***</td>
<td>1.939</td>
<td>6.606***</td>
<td>6.648***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(1.482)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Free-Riders (FR)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pun\textsuperscript{few}</td>
<td></td>
<td>2.166**</td>
<td>4.288***</td>
<td>3.643**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(0.975)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pun\textsuperscript{many}</td>
<td>4.469***</td>
<td>4.194***</td>
<td>2.500</td>
<td>1.225</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(0.988)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pun\text{it}</td>
<td>2.982***</td>
<td>2.794***</td>
<td>2.641***</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(0.599)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NPun\text{it}</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(0.487)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Constant</strong></td>
<td>7.906***</td>
<td>7.571***</td>
<td>6.376***</td>
<td>6.744***</td>
<td>8.860***</td>
<td>4.620</td>
</tr>
<tr>
<td>(1.332)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Obs.</strong></td>
<td>2,790</td>
<td>2,790</td>
<td>2,790</td>
<td>950</td>
<td>950</td>
<td>950</td>
</tr>
<tr>
<td><strong>R(^2)</strong></td>
<td>0.094</td>
<td>0.137</td>
<td>0.147</td>
<td>0.220</td>
<td>0.149</td>
<td>0.253</td>
</tr>
<tr>
<td><strong>AIC</strong></td>
<td>18248</td>
<td>18111</td>
<td>18082</td>
<td>940</td>
<td>949</td>
<td>949</td>
</tr>
</tbody>
</table>

*Notes:* Estimates from linear random-effects models for the \(R_p\)-game. Dependent variable: individual contribution per period. Dummies with superscript ‘few’ indicate that one, and dummies with ‘many’ indicate that two or three other subjects in the respective group are CC- or Pun-type. Columns (1)–(3) are based on the sample of conditional cooperators: \(N = 2,790\) \((279 CC\text{-types over 10 periods}); columns (4)–(6) use the sample of free-riders: \(N = 950\) \((95 FR\text{-types over 10 periods}). All specifications include a constant term and a full set of period-fixed dummies (coefficients not reported). Standard errors, clustered at the group level, are in parentheses; *** / ** / * indicate significance at the 1%--, 5%-,- and 10%-level, respectively.

the coefficients on the effect from having few or many Pun-types in a group remain quantitatively large and significant at the 1%- and 5%-level, respectively.\(^{29}\)

Estimations for the CC-types in the stranger design, which are presented in Table 5, show similar results. The \(CC\text{it}\) dummies are both insignificant (column 1), whereas the coefficients on the \(Pun\text{it}\) dummies are both large and relatively precisely estimated (column 2). When the two sets of dummies are combined, those for the prevalence of Pun-types in a matching-group remain highly significant. In addition, we can reject \(\lambda_1 = \lambda_3\) \((p = 0.008)\) and find borderline evidence when testing \(\lambda_2 = \lambda_4\) \((p = 0.1364)\).

Our results therefore show that — in the absence of group members who are willing to enforce a contribution norm — conditional cooperators per se do not necessarily perform well in

\(^{29}\)The Wald tests again reject \(\lambda_3 = \lambda_4\) \((p = 0.001)\). However, we do not detect a significant difference between the coefficients on CC\textsuperscript{many} and Pun\textsuperscript{many} \((p = 0.270)\).
coordinating on high contribution levels. Once pro-social punishers enter a group, conditional cooperators are much more willing to make higher contributions. The presence of $Pun$-types therefore seems to be essential for obtaining high contribution levels among conditional cooperators.

Next we turn to the results for free-riders. Overall, the estimates from columns (4)–(6) in Tables 4 and 5 provide a similar picture. However, due to the limited number of observations (we only observe 95 free-riders in the $R_p$, and 35 in the $R_s$ game), some of our findings are less instructive and somewhat under-powered. For the partner design, columns (4) and (5) of Table 4 suggest that $FR$-types’ contributions are, similar as those of $CC$-types, increasing in the number of $CC$- and $Pun$-types in their group. For the sample of free-riders, the coefficients on the $CC_{many}$ dummy becomes larger and is now significant at the 1% level (despite a larger standard error as compared to column (1)). Concerning the presence of pro-social punishers, we only find a large and statistically significant effect from having few (as compared to no) $Pun$-types. The $Pun_{many}$ dummy is insignificant (but not statistically different from $Pun_{few}$; testing $\lambda_3 = \lambda_4$ yields $p = 0.144$). Column (6), which presents the estimates for equation (6), suggests that the largest effect comes from having many $CC$-types in a group. Having more $Pun$-types further increases the free-riders’ contributions, but the effect is only statistically significant for one of the $Pun$-dummies.

From these estimates it appears tempting to conclude that the contributions of $FR$-types are more sensitive to the presence of conditional cooperators rather than pro-social punishers. However, a closer look at the data from the partner design shows an almost perfect overlap of $CC$- and $Pun$-types in the (few) groups of the free-riders.\textsuperscript{30} Hence, the high correlation among types in this small sample impedes our ability to draw strong conclusions on the differential impact of the two different types on free-riders’ behavior in the partner design.

For the stranger protocol (where the overlap of $CC$- and $Pun$-types in the matching groups is smaller), the results for the free-riders are much closer to those observed for the conditional cooperators. Columns (4) and (5) of Table 5 indicate that free-riders contribute significantly more, the more $CC$- and $Pun$-types are in their matching groups. For the model specification in column (6), the $CC_{\ell}$ dummies loose significance whereas the $Pun_{\ell}$ dummies remain large and highly significant. Post-estimation tests reject $\lambda_3 = \lambda_4$ ($p = 0.008$) as well as $\lambda_4 = \lambda_3$ ($p = 0.008$) and $\lambda_2 = \lambda_4$ ($p = 0.033$).

To wrap-up, the estimates show that free-riders’ contributions are influenced by both, the presence of $CC$- and $Pun$-types. While the data from the stranger protocol point to a clear enforcement result — a higher share of pro-social punishers in a matching group pushes free-riders to contribute more to the public good — the data from the partner protocol highlight the

\textsuperscript{30}In almost all cases when the $Pun_{many}$ dummy is equal to one, $CC_{many}$ is one, too.
Table 5: Group Composition and Individual Contributions (Stranger Design)

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Conditional Cooperators (CC)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CC\textsuperscript{few}</td>
<td>0.201</td>
<td>0.490</td>
<td>6.718***</td>
<td>-0.0456</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(2.398)</td>
<td>(2.313)</td>
<td>(2.360)</td>
<td>(0.530)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CC\textsuperscript{many}</td>
<td>2.747</td>
<td>2.648</td>
<td>11.17***</td>
<td>0.489</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(2.269)</td>
<td>(2.227)</td>
<td>(1.760)</td>
<td>(2.937)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>P\textsuperscript{few}</td>
<td>7.584***</td>
<td>7.302***</td>
<td>6.538***</td>
<td>6.413***</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(1.386)</td>
<td>(1.197)</td>
<td>(1.535)</td>
<td>(2.051)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>P\textsuperscript{many}</td>
<td>7.622***</td>
<td>6.837***</td>
<td>13.24***</td>
<td>12.82***</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(1.501)</td>
<td>(1.277)</td>
<td>(1.392)</td>
<td>(3.026)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>P\textsubscript{i}</td>
<td>2.568***</td>
<td>3.275***</td>
<td>3.053***</td>
<td>\</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.868)</td>
<td>(0.896)</td>
<td>(0.928)</td>
<td>\</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NP\textsubscript{i}</td>
<td>-2.863*</td>
<td>-4.079***</td>
<td>-4.014***</td>
<td>\</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(1.665)</td>
<td>(1.334)</td>
<td>(1.324)</td>
<td>\</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>8.398***</td>
<td>2.330*</td>
<td>1.387</td>
<td>1.216</td>
<td>1.691*</td>
<td>1.677*</td>
</tr>
<tr>
<td></td>
<td>(1.901)</td>
<td>(1.357)</td>
<td>(1.823)</td>
<td>(1.086)</td>
<td>(1.016)</td>
<td>(1.001)</td>
</tr>
<tr>
<td>Obs.</td>
<td>1,030</td>
<td>1,030</td>
<td>1,030</td>
<td>350</td>
<td>350</td>
<td>350</td>
</tr>
<tr>
<td>R\textsuperscript{2}</td>
<td>0.115</td>
<td>0.159</td>
<td>0.197</td>
<td>0.294</td>
<td>0.417</td>
<td>0.418</td>
</tr>
<tr>
<td>AIC</td>
<td>6428</td>
<td>6375</td>
<td>6332</td>
<td>2263</td>
<td>2184</td>
<td>2180</td>
</tr>
</tbody>
</table>

Notes: Estimates from linear random-effects models for the Rs-game. Dependent variable: individual contribution per period. Dummies with superscript ‘few’ indicate that two to four [three or four in columns 1–3], and dummies with ‘many’ indicate that five or more subjects in the respective matching group are CC- or Pun-type. (The pooling of dummies was based on the actual type allocation in the matching groups, with the objective to minimize loss of information.) Columns (1)–(3) are based on the sample of conditional cooperators: \( N = 1,030 \) (103 CC-types over 10 periods); columns (4)–(6) use the sample of free-riders: \( N = 350 \) (35 FR-types over 10 periods). All specifications include a constant term and a full set of period-fixed dummies (coefficients not reported). Standard errors, clustered at the group level, are in parentheses; *** / ** / * indicate significance at the 1%- , 5%- , and 10%-level, respectively.

Influence of conditional cooperators. While the latter observation is based on a small sample, it is consistent with the idea that (at least some) free-riders act strategically in the repeated game, playing high contributions that aim at encouraging reciprocal behavior of the CC-types (e.g., Sonnemans et al., 1999; Keser and van Winden, 2000; Muller et al., 2008).

A last point worth discussing is the fact that the estimates from Tables 4 and 5 allow for a comparison of the average contributions among the different types introduced in our type classification from above (see, e.g., Figure 3). To see this, one has to recognize that the constant term \( \lambda_0 \) from equation 6 captures a type’s average contribution. Focusing on the partner design, the estimates from column (3) therefore suggest that an average conditional cooperater, who is not classified as pro-social punisher (P\textsubscript{i} = 0), contributes 6.4 tokens (in the first period and with zero CC- and Pun-types among the group members). A CC × Pun-type, in contrast,
contributes significantly more: 9.0 tokens ($\lambda_0 + \phi$, based on column 3). From column (6) we further learn that an average free-rider, who is not classified as non-punisher ($NPun_i = 0$), makes a contribution of 4.6 tokens. A $FR \times NPun$-type would, cet. par., contribute significantly less: 1.1 tokens. The different cooperation patterns from the one-shot C-game as well as the heterogenous punishment patterns from the one-shot P-game (which are used to classify these different types) are therefore strong predictors of the sizeable differences in individual contribution levels that are observed for the repeated game.

5 Concluding Discussion

Using a parsimonious strategy-method approach, we presented systematic evidence on the heterogeneity of punishment patterns at the individual level. We linked our classification of punishment-types to the popular cooperation-type classification from Fischbacher et al. (2001). This allowed for an individual-level analysis of the relationship between subjects’ dispositions to cooperate and their inclinations to enforce cooperation via peer punishment. The resulting two-dimensional classification suggested the existence of two distinct behavioral archetypes. On the one hand, we identified many subjects whose punishment and cooperation patterns are aligned. On the other hand, our analysis uncovered a non-trivial fraction of subjects whose cooperation and punishment patterns diverged: free-riders that punished pro-socially and conditional co-operators that did not punish. Hence, for a majority of subjects cooperation and punishment indeed seem like two sides of the same coin. However, for a significant part of our sample, cooperation and punishment seem to be different behavioral traits.

The divergence between cooperation and punishment patterns further allowed us to assess the role of the two-dimensional variation in types — which we identified in two independent one-shot games — for explaining group outcomes and individual behavior in a third, repeated game with peer punishment. Our analyses provided strong, causal evidence on the relative importance of pro-social punishers for achieving and maintaining cooperation. While variation in cooperation types within a (matching) group explains large parts of the variance in group outcomes, similar variation in punishment types has a higher explanatory power.

The latter finding is relevant, since previous work has predominantly hinted at the importance of conditional cooperators for a group’s success (e.g., Gächter and Thöni, 2005; Burlando and Guala, 2005). Except for Rustagi et al. (2010), however, the corresponding inferences are usually drawn from situations that do not entail elements of punishment. Given that the absence of sanctioning opportunities in natural environments is likely to be the exception rather than the rule, actual group outcomes might not be determined by individuals’ cooperation types per se, but rather by the concomitant inclination to engage in pro-social punishment. Our results, in particular the identification of a behavioral archetype with diverging punishment and coop-
eration patterns, underline that this distinction indeed matters. It will be interesting to see in future studies if a similar differentiation also applies to other forms of pro- (e.g., Falk and Szech, 2013) and anti-social (e.g., Abbink and Serra, 2012) behavior.

The results and the methodologies from our study open several avenues for follow-up research. To advance our understanding of cross-cultural differences in cooperation (Henrich et al., 2006; Herrmann et al., 2008), one could readily apply our approach to examine the underlying variation in individual cooperation and punishment types. Exploring type variation in social dilemmas beyond linear public goods games (see, e.g., Cason and Gangadharan, 2015) will also help to reassess the underlying motivations of peer punishment. If, for instance, people solely punish to reduce inequality in payoffs (in a self-centered way, e.g., following Fehr and Schmidt, 1999) this could intuitively explain the aligned behavioral archetype (pro-socially punishing conditional cooperators as well as individuals who free-ride in both stages of the game). Depending on the parametrization of the game, self-centered models of inequality aversion might not be easily reconcilable with free-riders that are pro-social punishers or with conditional cooperators that do not punish. These diverging types would also be incompatible with a notion of strong reciprocity, assuming cooperation and punishment to be responses that are triggered by positive and negative reciprocity, respectively (Dohmen et al., 2008). Building on our design — e.g., by augmenting our strategy-method to account for a subject’s beliefs about others’ punishment — future research might address this point and disentangle the influence of rational motives (Casari and Luini, 2012), emotions (Falk et al., 2005; Reuben and van Winden, 2008; Hopfensitz and Reuben, 2009) or inconsistency (Blanco et al., 2011) in explaining the different archetypes and their punishment patterns.
References


Appendix

A1 Contribution Triples

Below we list the hypothetical contribution triples that were used within each of the ten combinations of $g^L$, $g^M$ and $g^H$ (see Section 2.1). Before the experiment, these $10 \times 8$ triples were randomly generated by sampling with replacement from the corresponding sets $g^L$, $g^M$, $g^H$. Each player then faced a randomly selected triple within each combination. If the selected triple would by chance correspond to the real triple, the subject would not face this situation; instead another one of the pre-defined contribution triples for the corresponding combination would be drawn.

\[
\begin{array}{cccccccc}
(1) & (2) & (3) & (4) & (5) & (6) & (7) & (8) \\
(g^L, g^L, g^L): & (0,0,0) & (0,2,3) & (1,1,3) & (1,2,2) & (1,2,3) & (1,2,4) & (1,3,3) & (1,3,4) \\
(g^L, g^L, g^M): & (0,1,5) & (0,2,8) & (0,2,14) & (1,2,10) & (1,2,12) & (1,3,14) & (2,2,6) & (2,3,12) \\
(g^L, g^L, g^H): & (0,3,18) & (1,2,20) & (1,3,19) & (1,4,20) & (2,2,18) & (2,3,14) & (3,2,11) & (3,3,17) \\
(g^L, g^M, g^M): & (0,9,11) & (0,5,12) & (0,13,14) & (1,10,15) & (2,6,8) & (2,9,11) & (2,10,15) & (3,13,14) \\
(g^L, g^M, g^H): & (0,6,19) & (0,14,17) & (2,6,17) & (2,8,20) & (2,11,19) & (3,7,18) & (4,8,17) & (4,10,20) \\
(g^L, g^H, g^H): & (0,18,19) & (1,19,19) & (2,18,19) & (2,18,20) & (2,19,19) & (3,18,20) & (3,19,19) & (4,19,20) \\
(g^M, g^M, g^M): & (5,7,12) & (5,14,15) & (6,6,9) & (6,10,10) & (7,8,9) & (7,10,13) & (7,14,15) & (8,9,11) \\
(g^M, g^M, g^H): & (5,5,17) & (5,8,16) & (6,11,20) & (8,15,17) & (9,12,18) & (9,15,18) & (11,15,19) & (12,15,19) \\
(g^M, g^H, g^H): & (5,18,20) & (7,18,19) & (9,18,20) & (11,17,17) & (12,17,18) & (12,18,18) & (14,17,20) & (15,17,19) \\
(g^H, g^H, g^H): & (17,17,19) & (17,18,19) & (17,18,20) & (17,19,19) & (17,19,20) & (18,18,19) & (18,18,20) & (20,20,20)
\end{array}
\]

A2 Type Distribution among (Matching-) Groups

Table A1 illustrates the group composition that emerged from the random assignment of subjects into different [matching-] groups. In addition, the table presents the expected distribution (numbers in italics) based on the population frequencies of CC- and Pun-types as reported in Tables 1 and 2, respectively. The chance, for instance, of having four CC-types in one group is given by $0.608^4$. Among 113 groups, one should thus expect 15.4 groups with this composition. Stated differently: the numbers in italics form the ‘perfect randomization’ benchmark. The actual outcome is in fact very close to this benchmark.

The top part of the table illustrates the variation in the different types among the 113 four-player groups in the partner protocol (R_p-game). Consistent with the high population frequency of conditional cooperators (60.8 % of our sample, see Table 2) we observe that the majority of groups are populated by two (35 groups) or three (48 groups) CC-types. In addition, there are several groups with no (4), one (13) or even four CC-types (13 groups). A slightly more symmetric distribution is observed for Pun-types — reflecting the fact that the population
Table A1: Type Distribution per (Matching) Group

<table>
<thead>
<tr>
<th>Number of subjects:</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>CC</td>
<td>4</td>
<td>13</td>
<td>35</td>
<td>48</td>
<td>13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>113</td>
</tr>
<tr>
<td>Pun</td>
<td>15</td>
<td>30</td>
<td>34</td>
<td>30</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>113</td>
</tr>
<tr>
<td>R_p-game</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CC</td>
<td></td>
<td>4.0</td>
<td>16.6</td>
<td>38.5</td>
<td>39.8</td>
<td>15.4</td>
<td></td>
<td></td>
<td></td>
<td>113</td>
</tr>
<tr>
<td>Pun</td>
<td></td>
<td>8.8</td>
<td>31.5</td>
<td>42.1</td>
<td>25.0</td>
<td>5.6</td>
<td></td>
<td></td>
<td></td>
<td>113</td>
</tr>
<tr>
<td>R_S-game</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CC</td>
<td></td>
<td>0.0</td>
<td>0.2</td>
<td>0.8</td>
<td>2.6</td>
<td>5.0</td>
<td>6.2</td>
<td>4.8</td>
<td>2.1</td>
<td>0.4</td>
</tr>
<tr>
<td>Pun</td>
<td></td>
<td>0.1</td>
<td>1.0</td>
<td>3.0</td>
<td>5.3</td>
<td>5.9</td>
<td>4.2</td>
<td>1.9</td>
<td>0.5</td>
<td>0.1</td>
</tr>
</tbody>
</table>

Notes: In the R_p-game subjects are counted at the group level (with 4 subjects per observational unit). In the R_S-game subjects are counted at the matching-group level (with 8 subjects per observational unit). The depicted distribution of subjects occurred from randomly assigning subjects to groups (matching-groups) at the beginning of the R-game. The numbers in italics present the expected distribution based on the population frequencies of CC- and Pun-types as reported in Tables 1 and 2, respectively.

prevalence is close to one half (47.1 % of our sample, see Table 1). There are between 30 to 34 groups, each with either one, two, or three Pun-types. In addition, there are 15 groups with zero and four groups with four Pun-types. We use two-sided Fisher’s exact tests to assess the hypothesis that the observed and the predicted distribution of groups with different type-compositions stem from the same distribution. Consistent with random group assignment, this H0 cannot be rejected (p = 0.812 for the distribution of CC-types, and p = 0.539 for the distribution of Pun-types).

The lower part of Table A1 captures the variation in group compositions between the 22 matching groups (each with eight subjects) from the stranger protocol (R_S-game). Similar as above, the data indicate quite some variation in the type composition across groups. Given the limited number of matching groups, there appear to be larger deviations from the expected number of groups with different compositions. However, the actual distribution is again not different from the expected random distribution: the p-values from two-sided Fisher’s exact tests are, exactly as above, p = 0.812 for the CC- and p = 0.539 for the Pun-types.
A3 Complementary Figures

Figure A1: Primary Classification Approach: Stylized Illustration of Punishment Types

<table>
<thead>
<tr>
<th></th>
<th>N\text{Pun}</th>
<th>\text{Pun}</th>
<th>A\text{Pun}</th>
<th>\text{NCL}</th>
</tr>
</thead>
<tbody>
<tr>
<td>$d_{ij}$</td>
<td>$d_{ij}$</td>
<td>$d_{ij}$</td>
<td>$d_{ij}$</td>
<td>$0$</td>
</tr>
<tr>
<td>$\hat{\alpha}_i = \hat{\beta}_i = 0$</td>
<td>$\hat{\beta}_i &gt; 0$ with $p \leq 0.01$</td>
<td>$\hat{\beta}_i &lt; 0$ with $p \leq 0.01$ or $\hat{\alpha}_i &gt; 0$ ($p \leq 0.01$) &amp; $\hat{\beta}_i$ insignif.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notes: Type classifications based on $\hat{\alpha}$ and $\hat{\beta}$ obtained from estimating eq. (3).

Figure A2: Self-Centered Classification Approach: Stylized Illustration of Punishment Types

<table>
<thead>
<tr>
<th></th>
<th>N\text{Pun}</th>
<th>\text{Pun}</th>
<th>A\text{Pun}</th>
<th>\text{NCL}</th>
</tr>
</thead>
<tbody>
<tr>
<td>$d_{ij}$</td>
<td>$d_{ij}$</td>
<td>$d_{ij}$</td>
<td>$d_{ij}$</td>
<td>$0$</td>
</tr>
<tr>
<td>$\hat{\beta}_1' = \hat{\beta}_2' &gt; 0$ with $p \leq 0.01$</td>
<td>$\hat{\beta}_1' = 0, \hat{\beta}_2' &gt; 0$ with $p \leq 0.01$</td>
<td>$\hat{\beta}_1' &gt; \hat{\beta}_2' &gt; 0$ with $p \leq 0.01$</td>
<td>$\hat{\beta}_1' &gt; \hat{\beta}_2' &gt; 0$ with $p \leq 0.01$</td>
<td></td>
</tr>
<tr>
<td>$\hat{\beta}_1' &lt; 0, \hat{\beta}_2' &gt; 0$ with $p \leq 0.01$</td>
<td>$\hat{\beta}_1' &lt; 0, \hat{\beta}_2' &gt; 0$ with $p \leq 0.01$</td>
<td>$\hat{\beta}_1' &lt; 0, \hat{\beta}_2' &gt; 0$ with $p \leq 0.01$</td>
<td>$\hat{\beta}_1' &gt; 0, \hat{\beta}_2' &gt; 0$ with $p \leq 0.01$</td>
<td></td>
</tr>
</tbody>
</table>

Notes: Different types based on $\hat{\beta}_1'$ and $\hat{\beta}_2'$ from estimating eq. (4). Panels a to d show different patterns of pro-socially punishing subjects. Panels a.i to d.i display anti-social patterns. The self-centered pattern b would be classified as S\text{Pun}', whereas pattern a would be labeled S\text{Pun}'. As long as $\hat{\beta}_1'$ is not significantly positive, pattern c would be classified as S\text{Pun}', too. Empirically, we hardly observe patterns as those from panels c and d. Patterns e and e.i are mechanically possible but are not observed in our sample. (Patterns for N\text{Pun}'- or N\text{CL}'-types are not illustrated here.)
Figure A3: Distribution of $\hat{\beta}_i$ and $\hat{b}_i$

Notes: Scatter plot for individual level peer punishment pattern slope $\hat{\beta}_i$ and contribution pattern slope $\hat{b}_i$ for the four most prevalent types, i.e., CC, FR, Pun, and NPun. The estimated correlation between the respective $\hat{\beta}_i$ and $\hat{b}_i$ is depicted as a yellow line. To ease illustration $FR \times NPun$-type values are not plotted. The concentration of observations at $\hat{b}_i = 0$ and $\hat{b}_i = 1$ is due to ‘perfect’ free-riders and ‘perfect’ conditional cooperation, respectively. The former never contribute in the C-game, whereas the latter types perfectly (i.e., 1:1) match the average group contribution.