ECONtribute
MARKETS & PUBLIC POLICY

ECONtribute
Discussion Paper No. 323

Maintaining Cooperation through Vertical Communication of Trust when Removing Sanctions

Ann-Christin Posten
Sebastian Kube
Pınar Uğurlar
Joris Lammers

July 2024
www.econtribute.de

UNIVERSITÄT BONN
Universität zu Köln

Funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2126/1-390838866 is gratefully acknowledged.

"Maintaining Cooperation through Vertical Communication of Trust when Removing Sanctions"

Ann-Christin Posten[1, 2]

Pınar Uğurlar[3]

Sebastian Kube[4,5,6]

Joris Lammers[7]

[1]Department of Psychology, University of Limerick, Limerick, Ireland

[2]Center for Social Issues – Research, University of Limerick, Limerick, Ireland

[3]Department of Psychology, Özyeğin University, Istanbul, Turkey

[4]Institute for Applied Microeconomics, University of Bonn, Bonn, Germany

[5]Center for Economics and Neuroscience, University of Bonn, Germany

[6]Max Planck Institute for Research on Collective Goods, Bonn, Germany

[7]Social Cognition Center Cologne, University of Cologne, Cologne, Germany

Address Correspondence to: Ann-Christin Posten, Department of Psychology, University of Limerick, Limerick, V94 T9PX Ireland, ann-christin.posten@ul.ie

Abstract

An effective way to foster cooperation is to monitor behaviour and sanction freeriding. Yet, previous studies have shown that cooperation quickly declines when sanctioning mechanisms are removed. We test if explicitly expressing trust in players' capability to maintain cooperation after the removal of sanctions, i.e. vertical communication of trust, has the potential to alleviate this drop in compliance. Four incentivized public-goods experiments (N = 2423) find that the vertical communication of trust maintains cooperation upon the removal of centralized (Study 1), third-party (Study 2), and peer punishment (Study 3), and this effect extends beyond single interactions (Study 4). In all studies, vertical trust communication increases mutual trust among players, providing support to the idea that vertically communicating trust can be a self-fulfilling prophecy. Extrapolating our findings to natural environments, they suggest that authorities should carefully consider how they communicate the lifting of rules and sanctions.

Keywords: Cooperation, Vertical Trust, Punishment, Public Good, Experiment

Maintaining Cooperation through Vertical Communication of Trust when Removing Sanctions

Modern social and environmental challenges are often rooted in cooperation problems. These cooperation problems emerge due to a conflict between individual rationality and the collective perspective[1]. Among the most popular mechanisms to mitigate such cooperation problems is monitoring with various forms of sanctioning institutions. For example, in response to the COVID-19 pandemic, many countries introduced sanctions to enforce compliance with cooperative public health measures. Yet, when the need for these measures decreased—because of dropping infection rates and increasing vaccination rates—political pressure put an end to the sanctioning mechanisms and selfish behaviour increased[2]. Although it is in the collective interest to voluntarily continue to cooperate, this detrimental effect on cooperation rates after the removal of sanctions has been frequently shown in lab studies[3–6]. This underscores the importance of finding ways to sustain cooperation, even after the removal of sanctioning mechanisms. In this paper, we thus ask if communication can fulfil this purpose - not only in the context of centralized punishment (as in the COVID-19 example), but also more generally for punishment by a (otherwise unrelated) third-party and mutual second-party (peer) punishment.

Our specific focus lies on the (unidirectional) vertical communication from the principal, the authority responsible for removing the punishment mechanism, to the agents encountering the cooperation problem. More precisely, we empirically test whether cooperation can be maintained if the relevant authorities explain – when they communicate the removal of the punishment institution – that they do so because they trust the population to cooperate in its absence. By investigating the efficacy of *vertical* communication (from the principal to the population), we complement the prevalent focus in existing literature on horizontal

communication (among the members of the population) as a substitute for punishment to foster cooperation[7,8]. We expect vertical communication to be an effective instrument for maintaining cooperation when sanctioning mechanisms are removed, because providing the explanation of trust driving the change might act as a self-fulfilling prophecy[9]: the vertical communication of trusting the population to cooperate shapes the beliefs within the population that others can still be trusted, which in turn sustains their willingness to cooperate, thus maintaining cooperation levels and showing that others can indeed be trusted to cooperate.

This line of reasoning is similar to an equilibrium-selection idea for societies harbouring multiple equilibria. The idea is proposed by Dasgupta[10], for example, who also stresses the importance of mutual trust as a basis for cooperation. Indeed, Thöni et al.[11] observe a positive correlation between trust, as measured by standard trust questions from the World Values Survey, and cooperation behaviour in an incentivized cooperation game. This link is also found by Kim et al.[12] when combining lab experimental data from trust and cooperation games. They show that people infer cooperativeness from trustworthiness and cooperate more when they predict others to be more cooperative. Similarly, based on survey data, Schmelz[13] stresses the essential role of trust in government for citizens' voluntary and non-enforced cooperation with COVID-19 policies. Complementing these behavioural findings on trust and cooperation, simulation results reported in Battu and Rahwan[14] suggest that "society should focus on creating a critical amount of trust to harness the conditional nature of its members".

Vertical communication is of particular significance, given the importance of conditionality and beliefs about others cooperativeness for shaping subsequent cooperative behaviour. If a sizable fraction of a population acts like conditional cooperators[15], the beliefs that people hold about one another become key to cooperation[16,17] and shaping beliefs through

vertical communication should have significant effects on subsequent behaviour. Consistent with this idea, Engel et al.[18] find that groups cooperate more after being provided with selective positive information about the cooperative behaviour of other, unrelated groups from the same population. Vice versa, Galbiati et al.[19] observe that sanctions can be perceived as a negative signal about others' cooperativeness, if they are actively installed by an informed third party, and thus make cooperation in a minimum-effort game more difficult. Relatedly, but focusing on horizontal communication, Tyran and Feld[20], Gürerk et al.[5], and Sutter et al.[21] all demonstrate that endogenous institutional choices, made by the population via voting, can serve as a signal of others' cooperativeness.

To provide causal evidence for the effectiveness of vertical communication on cooperation behaviour, we employ an incentivized public-goods game (PGG), a workhorse frequently used in the literature to study cooperation problems. Participants form groups and individually decide how many tokens from their private endowment they want to contribute to a common pool. This pool is then multiplied and shared equally among all group members. The game parameters are chosen such that the group payoff is maximized if all tokens are contributed, but selfish money-maximizing behaviour leads to the equilibrium of zero contributions. In a first round, participants play the PGG with a sanctioning institution in place, either a centralized punishment scheme in Study 1, third-party punishment in Study 2, or peer punishment in Study 3. This is then followed in a second round by the regular PGG without any sanctioning institution. In each study, conditions vary by how the removal of sanctions between the first and second round is communicated to the participants. In addition to using contributions as a proxy for cooperation behaviour, trust is assessed by using a standard Likert-type question. For all three sanctioning institutions, we find that the vertical communication of trust affects self-

reported trust and helps to maintain cooperation despite the removal of the sanctioning mechanism. Upon vertical trust communication, across all studies, contributions were on average over 25% higher than without communicating a reason for sanctioning removal.

Method and Results

Four preregistered, incentivized studies investigated the effectiveness of the vertical communication of trust. Participants could only take part in one of the four studies. The studies were run online using Qualtrics. Participants were US-citizens that were randomly sampled from the participant pool of Prolific, a UK-based service provider specialized in conducting academic research. Payments on prolific are generally made in GBP, so we used British Pence throughout our studies, paying GBP 0.10 per 10 tokens earned in the game (in addition to a base payment). Studies were conducted consistent with ethical principles provided by the German Psychological Society DGPS[22] and exempt from Institutional Review Board approval following these same guidelines. Data, code, and materials are available at osf.io/duvrc/?view_only=1d4dcde1ac0d4970aae4ce432f110640 (18). Studies were preregistered at aspredicted.org (Study 1: aspredicted.org/LN2_5V2; Study 2: aspredicted.org/HB7_WGG; Study 3: aspredicted.org/5PJ_1FM; Study 4: aspredicted.org/NTY_1PQ).

The basic underlying game in each study was a four-player public-goods game, a game that is commonly used in the literature to study cooperation problems. Each round, players decided how to distribute 10 tokens between an individual and a collective investment. All tokens invested individually were tripled and returned only to the investing player, while all tokens invested collectively were doubled and the doubled amount was paid out to each of the four players in the group (following a similar parametrization as used, for instance, in Andreoni and Gee 2012[23]). The individual payoff-maximizing strategy is to keep own tokens in the

individual investment pot, independent of what the others (are expected to) do, in which case each player would earn 10*3 = 30 tokens from the respective private investment. The cooperative outcome is to invest the entire endowment in the collective pot, in which case each player would earn a return of (4*10)*2 = 80 tokens from the public investments. The maximum payoff per round for an individual player was 90 tokens if the player only invested individually and the other 3 players only invested in the collective (i.e., 10*3 tokens from the private investment and (3*10)*2 = 60 tokens from the other 3 players' public investments). The minimum payoff was 20 tokens if the player only invested in the public good and all other players only invested individually (i.e., 10*2 = 20 tokens from the public investment and (3*0)*2 = 0 tokens from the other 3 players' missing public investment).

Each study consisted of a first round where the PGG was accompanied by a sanctioning institution, followed by a second round where the sanctions were removed. The three sanctioning institutions that we used are the most prominent ones in the literature on cooperation and punishment and closely follow the respective parametrizations[23]. At the beginning, all participants were informed about the total number of rounds they were about to play, and sanctioning parameters were explained. The removal of sanctions was communicated in different ways, depending on the treatment condition. Players received feedback on others' contributions, received punishment and realized payoffs only at the very end of the study. The setup was completely anonymous, and the groups were randomly drawn each round.

Study 1

Our first study (n = 604) investigates if the vertical communication of trust can maintain cooperation after a *centralized* sanctioning mechanism is removed. A centralized mechanism closely mirrors sanctioning mechanisms imposed by authorities, similar to our opening example

on COVID-19. The obligation under this mechanism was to invest the socially optimal amount into the public investment. If a player was detected to deviate from this obligation, that person had to pay a non-deterrent fine depending on the size of the deviation. This was implemented as follows: one participant of each group was randomly drawn and punished via the subtraction of 1.6 tokens for each token that this player had not invested in the collective (see, for instance, Galbiati and Vertova[24] for a similar approach). In the second round, the sanctioning mechanism was removed, and the vertical communication took place. When introducing the removal, participants were informed that this happened because the researchers "trust that most people invest sufficiently" (*trust condition*), or that the researchers "do not trust that most people invest sufficiently" (*distrust condition*), or participants did not receive any explanation (*control condition*). Upon their decision in round 2, participants indicated their trust in their co-players and indicated further affective states and a manipulation check (see supplementary information for further study descriptions and analyses).

Our main variable of interest is the contributed amount as the proxy for cooperation. Results showed that vertical communication of trust kept contributions at a high level upon sanctioning removal, while contributions in the control and distrust condition strongly dropped in their absence (interaction effect of $F(2, 601) = 3.26$, $p = .039$, $\eta^2 = .011$). This resulted in significantly higher contributions in the trust condition ($M = 5.91$, $SD = 3.79$) as compared to the distrust ($M = 4.87$, $SD = 4.02$), $t(601) = 2.68$, $p = .007$, $d = 0.27$, and control condition ($M = 4.99$, $SD = 3.92$), $t(601) = 2.36$, $p = .018$, $d = 0.24$, with the latter two not differing significantly, $t(601) = 0.32$, $p = .750$. This pattern supports the prediction that people by default expect others to defect upon removal of a sanctioning mechanism and that therefore cooperation drops – but that vertical communication of trust in the ability to contribute even in the absence of sanctions

prevents that drop in cooperation. In fact, shedding more light on the underlying psychological process by a mediation analysis suggests that trust in co-players, assessed by the seven-point Likert-type question “Do you think the other players could be trusted?”, acted as a critical psychological process-variable, mediating the effect of trust and distrust on cooperation (Fig. 1a; see Supplementary Information for full mediation analysis results).

Study 2

Our second study (n = 606) investigates if the vertical communication of trust can maintain cooperation after a *third-party* sanctioning mechanism is removed. This is comparable to judicial systems across the world where court judges have some discretion over the punishment (see, for instance, Fehr and Fischbacher[25]). Different from the centralized institution in Study 1, where obligation and fines are known in advance and carried out automatically, both the basis for and the size of the punishment are now decided by an actual player outside the group. Therefore, at the very end of the study one randomly selected player of each group was given the opportunity to punish the players of another group by subtracting up to 15 tokens (which were destroyed) from their individual payoffs. The third party was able to tailor the size of the punishment to the contribution levels, using the strategy method approach.

After the first round, the punishment mechanism was revoked with the same vertical communication as before, again creating a trust, control, and distrust condition. As in Study 1, an interaction effect confirmed, $F(2, 603) = 13.52$, $p < .001$, $\eta^2 = .04$, that, upon removal of the sanctioning mechanism, vertically communicating trust in people kept contributions at a significantly higher level (M = 6.43, SD = 3.74) compared to when communicating distrust (M = 5.22, SD = 3.86), $t(603) = 3.20$, p = .001, d = 0.32, or without communication in the control condition (M = 4.83, SD = 3.81), $t(603) = 4.24$, $p < .001$, d = 0.42. No difference between the

distrust and control conditions emerged, $t(603) = 1.04$, $p = .301$. As in Study 1, trust in others acted as a mediating process variable (Fig. 1b; see Supplementary Information for full mediation analysis results).

Study 3

Aside from formal institutions and their agents also fellow citizens and peers can punish non-cooperative behaviour. Study 3 ($n = 610$) mimics this peer-punishment by employing a *peer* punishment mechanism, as it is frequently used in existing literature[4]. This time, one randomly drawn player per group had the opportunity to punish the members of the same group. Again, punishment decisions were provided at the end of the study via the strategy method. Results show that vertically communicating trust was successful in maintaining high contributions, while contributions in the control and distrust condition strongly dropped in their absence (interaction effect: $F(2, 607) = 14.37$, $p < .001$, $\eta^2 = .05$), resulting in significantly lower contributions in the distrust ($M = 5.29$, $SD = 3.62$), $t(607) = 3.35$, $p = .001$, $d = 0.34$, and control condition ($M = 5.02$, $SD = 3.81$), $t(607) = 4.11$, $p < .001$, $d = 0.41$, compared to the trust condition ($M = 6.49$, $SD = 3.35$). Again, there was no difference between the distrust and control condition, $t(607) = 0.77$, $p = .442$. Trust in others again acted as a mediating process variable (Fig. 1c; see Supplementary Information for full mediation analysis results).

Study 4

In Studies 1-3, we always compared the effect of vertical trust communication (versus distrust or control) while removing a sanction on contributions in a second round. Our final study ($n = 603$) addressed two limitations. First, we sought to directly compare the effect of vertical communication of trust with that of maintaining a sanctioning mechanism. Second, we sought to test if the effect of vertical communication of trust maintains over time. Therefore, we replicated

Study 3 and extended the design to three rounds of public goods games. In the first round, the peer punishment mechanism was in place for all participants. In round 2, we compared three conditions: We used the same trust and distrust manipulations as before, but this time also included a third condition, in which the sanctioning mechanism was maintained in round 2 (i.e., nothing changed compared to round 1). This enabled us to examine how effective the vertical trust communication is compared to maintaining the sanctioning institution. In the third round, the sanctioning mechanism was removed in the third condition as well, without providing a reason for the removal. In the trust and distrust conditions the sanctioning institution remained absent in the third round, so all participants played the public goods game in the third round without a sanctioning institution being in place. Inspecting cooperation in round 2 showed that the trust condition (M = 6.10, SD = 3.85) was comparably high as the third condition with the ongoing sanctioning mechanism (M = 6.34, SD = 3.34) and did not differ, $t(600) = 0.67$, p = .505, suggesting that vertical communication of trust when removing sanctions is similarly effective as maintaining the sanctioning mechanism (Fig. 2). The distrust condition (M = 4.73, SD = 3.59) dropped in comparison to both the trust condition, $t(600) = 3.81$, $p < .001$, $d = 0.37$, and the maintained-sanctioning condition, $t(600) = 4.47$, $p < .001$, $d = 0.46$. Results of round 3 (in which none of the conditions featured a sanctioning mechanism) showed the clear advantage of the previous vertical communication of trust when removing sanctioning: although the round 3 instructions were identical across all three conditions, the mere prior vertical communication of trust in round 2 ensured that contributions in round 3 were higher in the trust condition (M = 5.66, SD = 3.83) than in the distrust condition (M = 4.87, SD = 3.62), $t(600) = 2.09$, p = .037, d = 0.21, and the (third) control condition (M = 4.78, SD = 3.96) $t(600) = 2.29$, p = .023, $d = 0.23$. The latter two did not differ, $t(600) = 0.19$, p = .846.

Discussion

Sanctioning institutions have been shown to effectively foster cooperation[26]. However, reasons might arise that lead to the institutions being removed, such as diminishing political support or increasing implementation cost. In fact, both Markussen et al.[27] and Kamei et al.[28] demonstrate that, in the case of a cooperation problem, as the cost of a sanctioning institution increases, participants tend to vote for the implementation of alternative measures that are less costly. Also, the alternative of keeping up the obligations without backing it up by sanctions for misbehaviour is usually only of limited efficacy[24]. Yet, maintaining cooperation in the absence of sanctioning institutions is of great societal importance. In this paper, we explored if the vertical communication of trust from authorities to the population is able to maintain cooperation. For three popular sanctioning institutions, centralized punishment, peer punishment and third-party punishment, we found that one can maintain cooperation, despite the removal of sanctions, by communicating trust in people's continued cooperation (Studies 1-3). This act of communication served as a self-fulfilling prophecy, enabling cooperation to be sustained at a level similar to when sanctions were still in place (Study 4). Additionally, the positive consequences of this top-down trust-based approach persisted beyond one immediate decision to contribute, whereas the effects of sanctions vanished immediately in their absence (Study 4).

We measured mood and other affective states[29] in relation to the last round of the PGG, in which our critical differences between trust, control, and distrust condition occurred. In all four studies, results suggested that mood did not account for the observed effects because, although we found in some studies significant differences between treatment conditions in empowerment, happiness, and/or surprise, the pattern of results never mirrored those on

cooperation (see Table 1). This is consistent with other work suggesting that mood does not have strong direct effects on cooperative behaviour[30].

Our findings add to the central importance of trust in cooperation[31]. The removal of sanctions undermines any potential default trust that people may have held before[32–34]. It leads to the quick inference that others may be untrustworthy and cause a self-fulfilling prophecy of defection. We show that a brief vertical communication of trust prevents this and instils a self-fulfilling prophecy of trust that maintains compliance even in the absence of enforcement. These insights are interesting for researchers working on conditional cooperation, because they add substantially to established findings on the importance of beliefs about others' cooperation and demonstrate an alternative method for shaping these beliefs. Moreover, our mediation analyses suggest that the same psychological process-variable, self-reported trust in other group members, is critical in all three sanctioning institutions. These insights are interesting for practitioners, because they identify vertical communication as an easy and effective way of maintaining cooperation. In particular when cooperative behaviour has developed over a longer timespan under a sanctioning institution that is about to be abolished, communicating trust and confidence in people's ability to act in a considerate manner not only appears quite effective but also particularly cost-effective.

Follow-up research could examine how vertical communication of trust compares to the communication of different content or by different senders, e.g., senders with or without stakes in the underlying cooperation game, or in-group compared to out-group senders. The communication of injunctive or descriptive norms might be of particular relevance[35,36], since the authority’s vertical trust communication in our study might also have shifted the focus to norms of cooperation[37,38]. In future research, the comparison of our results to conditions where norms

are explicitly communicated would allow to disentangle the focusing and informational effect of norms[39] from the effect of inducing trust and shaping beliefs for conditional cooperators. Measuring norms explicitly could also contribute to our understanding whether removal of sanctioning shifts the perceived relational model and its corresponding norms on cooperation, as suggested by Relational Models Theory[40]. Moreover, it would be interesting to see in future research whether the beneficial effects of vertical trust communication extend to other settings in which cooperation is critical, such as organizations or educational institutions.

Acknowledgements

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2126/1-390838866 and the research grant PO 1850/3-1 awarded to Ann-Christin Posten. Declaration of AI and AI-assisted technologies in the writing process: During the preparation of this work the authors used you.com in order to improve readability and language. After using this service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

1. Fehr, E. & Schurtenberger, I. Normative foundations of human cooperation. *Nat. Hum. Behav.* **2**, 458–468 (2018).
2. Yong, J. C. & Choy, B. K. C. Noncompliance with safety guidelines as a free-riding strategy: An evolutionary game-theoretic approach to cooperation during the COVID-19 pandemic. *Front. Psychol.* **12**, 646892 (2021).
3. Fehr, E. & Gächter, S. Cooperation and punishment in public goods experiments. *Am. Econ. Rev.* **90**, 980–994 (2000).
4. Fehr, E. & Gächter, S. Altruistic punishment in humans. *Nature* **450**, 137–140 (2002).
5. Gürerk, Ö., Irlenbusch, B. & Rockenbach, B. The competitive advantage of sanctioning institutions. *Science* **312**, 108–111 (2006).
6. Herrmann, B., Thöni, C. & Gächter, S. Antisocial punishment across societies. *Science* **319**, 1362–1367 (2008).
7. Bochet, O., Page, T. & Putterman, L. Communication and punishment in voluntary contribution experiments. *J. Econ. Behav. Organ.* **60**, 11–26 (2006).
8. Cardenas, J. C., Stranlund, J. & Willis, C. Local environmental control and institutional crowding-out. *World Dev.* **28**, 1719–1733 (2000).
9. Bacharach, M., Guerra, G. & Zizzo, D. J. The self-fulfilling property of trust: An experimental study. *Theory Decis.* **63**, 349–388 (2007).
10. Dasgupta, P. Trust and cooperation among economic agents. *Philos. Trans. R. Soc. B Biol. Sci.* **364**, 3301–3309 (2009).

11. Herrmann, B., Thöni, C. & Gächter, S. Antisocial punishment across societies. *Science* **319**, 1362–1367 (2008).
12. Kim, J., Putterman, L. & Zhang, X. Trust, beliefs and cooperation: Excavating a foundation of strong economies. *Eur. Econ. Rev.* **147**, 104166 (2022).
13. Schmelz, K. Enforcement may crowd out voluntary support for COVID-19 policies, especially where trust in government is weak and in a liberal society. *Proc. Natl. Acad. Sci.* **118**, e2016385118 (2021).
14. Battu, B. & Rahwan, T. Cooperation without punishment. *Sci. Rep.* **13**, 1213 (2023).
15. Fischbacher, U., Gachter, S. & Fehr, E. Are people conditionally cooperative? Evidence from a public goods experiment. *Econ. Lett.* (2001).
16. Fischbacher, U. & Gächter, S. Social preferences, beliefs, and the dynamics of free riding in public goods experiments. *Am. Econ. Rev.* **100**, 541–556 (2010).
17. Gächter, S., Kölle, F. & Quercia, S. Reciprocity and the tragedies of maintaining and providing the commons. *Nat. Hum. Behav.* **1**, 650–656 (2017).
18. Engel, C., Kube, S. & Kurschilgen, M. Managing expectations: How selective information affects cooperation and punishment in social dilemma games. *J. Econ. Behav. Organ.* **187**, 111–136 (2021).
19. Galbiati, R., Schlag, K. H. & Van Der Weele, J. J. Sanctions that signal: An experiment. *J. Econ. Behav. Organ.* **94**, 34–51 (2013).
20. Tyran, J. & Feld, L. P. Achieving compliance when legal sanctions are non-deterrent. *Scand. J. Econ.* **108**, 135–156 (2006).
21. Sutter, M., Haigner, S. & Kocher, M. G. Choosing the carrot or the stick? Endogenous institutional choice in social dilemma situations. *Rev. Econ. Stud.* **77**, 1540–1566 (2010).

22. Deutsche Gesellschaft für Psychologie (DGPs). *Ethisches Handeln in Der Psychologischen Forschung Empfehlungen Der Deutschen Gesellschaft Für Psychologie Für Forschende Und Ethikkommissionen*. vol. 1 (Hogrefe, Göttingen, Germany, 2018).
23. Andreoni, J. & Gee, L. K. Gun for hire: Delegated enforcement and peer punishment in public goods provision. *J. Public Econ.* **96**, 1036–1046 (2012).
24. Galbiati, R. & Vertova, P. Obligations and cooperative behaviour in public good games. *Games Econ. Behav.* **64**, 146–170 (2008).
25. Fehr, E. & Fischbacher, U. Third-party punishment and social norms. *Evol. Hum. Behav.* **25**, 63–87 (2004).
26. Balliet, D., Mulder, L. B. & Van Lange, P. A. M. Reward, punishment, and cooperation: A meta-analysis. *Psychol. Bull.* **137**, 594–615 (2011).
27. Markussen, T., Putterman, L. & Tyran, J.-R. Self-organization for collective action: An experimental study of voting on sanction regimes. *Rev. Econ. Stud.* **81**, 301–324 (2014).
28. Kamei, K., Putterman, L. & Tyran, J.-R. State or nature? Endogenous formal versus informal sanctions in the voluntary provision of public goods. *Exp. Econ.* **18**, 38–65 (2015).
29. Schwarz, N. & Clore, G. L. Mood, misattribution, and judgments of well-being: Informative and directive functions of affective states. *J. Pers. Soc. Psychol.* **45**, 513–523 (1983).
30. Hertel, G., Neuhof, J., Theuer, T. & Kerr, N. L. Mood effects on cooperation in small groups: Does positive mood simply lead to more cooperation? *Cogn. Emot.* **14**, 441–472 (2000).
31. Balliet, D. & Van Lange, P. A. M. Trust, conflict, and cooperation: A meta-analysis. *Psychol. Bull.* **139**, 1090–1112 (2013).

32. Schul, Y., Mayo, R. & Burnstein, E. Encoding under trust and distrust: The spontaneous activation of incongruent cognitions. *J. Pers. Soc. Psychol.* **86**, 668–679 (2004).

33. Schul, Y., Mayo, R. & Burnstein, E. The value of distrust. *J. Exp. Soc. Psychol.* **44**, 1293–1302 (2008).

34. Katzir, M. & Posten, A.-C. Are there dominant response tendencies for social reactions? Trust trumps mistrust—evidence from a Dominant Behavior Measure (DBM). *J. Pers. Soc. Psychol.* **125**, 57–81 (2023).

35. Cialdini, R. B., Reno, R. R. & Kallgren, C. A. A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places. *J. Pers. Soc. Psychol.* **58**, 1015 (1990).

36. Keizer, K., Lindenberg, S. & Steg, L. The importance of demonstratively restoring order. *PLoS ONE* **8**, e65137 (2013).

37. Gavrilets, S. & Richerson, P. J. Collective action and the evolution of social norm internalization. *Proc. Natl. Acad. Sci.* **114**, 6068–6073 (2017).

38. Blair, M. M. & Stout, L. A. Trust, trustworthiness, and the behavioral foundations of corporate law. *Univ. Pa. Law Rev.* **149**, 1735 (2001).

39. Krupka, E. & Weber, R. A. The focusing and informational effects of norms on pro-social behavior. *J. Econ. Psychol.* **30**, 307–320 (2009).

40. Fiske, A. P. The four elementary forms of sociality: Framework for a unified theory of social relations. *Psychol. Rev.* **99**, 689–723 (1992).

Figures

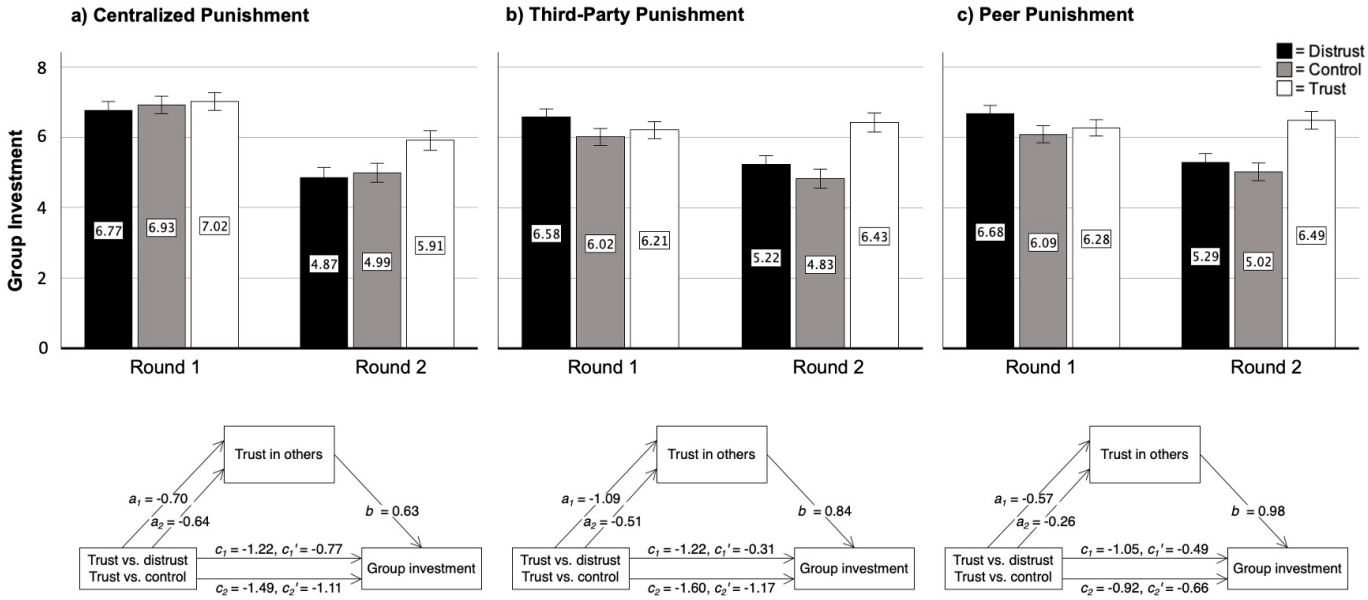

Fig. 1. Studies 1-3: Upon removal of sanctioning mechanisms in round 2, the vertical communication of trust as reason for the removal lead to more cooperation (group investment), measured as tokens contributed to a public good, than no communication about a reason (control condition) or communicating distrust (distrust condition). This was true for centralized punishment in Study 1 (a), third-party punishment in Study 2 (b), and peer punishment in Study 3 (c). Trust in others meditated the relationship between vertical trust communication and group investment comparing the trust with the distrust condition (path a_1) in all three studies and with the control condition (path a_2) in Studies 2-3. Error bars represent +/- 1 SE.

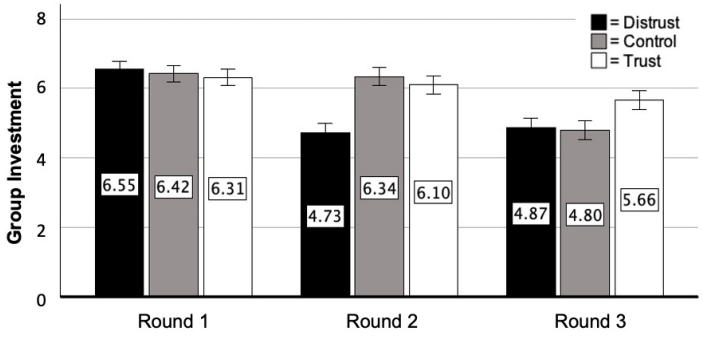

Fig. 2. Study 4: Removing sanctioning mechanisms in round 2 together with vertical trust communication (trust condition) led to no significant difference in contributions to the public good compared to the control condition, where sanctions were still installed, and to higher contributions than if distrust was communicated (distrust condition). In round 3, when punishment institutions where absent in all three conditions, having signaled trust prior to round 2 yielded higher contributions in round 3 compared to the other two conditions in which no trust had been communicated. Error bars represent +/- 1 SE.

Table 1

Means (and Standard Deviations) of affective State Ratings for Studies 1-4.

Variable	Condition	Study 1	Study 2	Study 3	Study 4
Empowered	Trust	2.71 *(1.28)*[b]	2.86 *(1.23)*[a]	2.57 *(1.19)*	2.97 *(1.14)*
	Control	3.03 *(1.28)*[c]	2.63 *(1.24)*	2.70 *(1.26)*[c]	2.79 *(1.25)*
	Distrust	2.68 *(1.38)*	2.41 *(1.16)*	2.37 *(1.23)*	2.73 *(1.24)*
Competent	Trust	3.30 *(1.19)*	3.36 *(1.18)*	3.22 *(1.12)*	3.31 *(1.17)*
	Control	3.36 *(1.18)*	3.27 *(1.23)*	3.19 *(1.11)*	3.27 *(1.19)*
	Distrust	3.20 *(1.22)*	3.14 *(1.17)*	3.06 *(1.12)*	3.22 *(1.18)*
Incompetent	Trust	1.36 *(0.73)*	1.33 *(0.76)*	1.37 *(0.80)*	1.44 *(0.86)*
	Control	1.43 *(0.93)*	1.32 *(0.72)*	1.30 *(0.67)*	1.33 *(0.73)*
	Distrust	1.36 *(0.77)*	1.34 *(0.74)*	1.38 *(1.16)*	1.46 *(0.86)*
Happy	Trust	2.78 *(1.16)*	2.76 *(1.20)*[a]	2.74 *(1.13)*[a]	2.82 *(1.14)*
	Control	2.89 *(1.29)*[c]	2.70 *(1.11)*[c]	2.69 *(1.11)*[c]	2.87 *(1.12)*
	Distrust	2.58 *(1.36)*	2.29 *(1.12)*	2.33 *(1.06)*	2.80 *(1.15)*
Angry	Trust	1.17 *(0.59)*	1.15 *(0.47)*	1.16 *(0.48)*	1.17 *(0.53)*
	Control	1.17 *(0.55)*	1.18 *(0.58)*	1.17 *(0.51)*	1.19 *(0.60)*
	Distrust	1.25 *(0.72)*	1.23 *(0.59)*	1.28 *(0.66)*	1.22 *(0.54)*
Surprised	Trust	1.92 *(1.16)*	1.85 *(1.03)*	1.87 *(1.05)*	1.77 *(1.01)*
	Control	2.02 *(1.21)*	1.69 *(0.95)*	1.70 *(1.08)*[c]	1.86 *(1.14)*
	Distrust	1.25 *(0.72)*	1.87 *(1.01)*	1.97 *(1.05)*	1.68 *(1.01)*

Note. Affective states were measured with respect to round 2 in Studies 1-3 and round 3 in Study 4. Superscripts indicate significant contrasts at the level of $p < .05$ between individual experimental condition: *a* indicates a significant contrast between the trust and the distrust condition, *b* indicates a significant contrast between the trust and the control condition, and c indicates a significant contrast between the control and distrust condition. Test statistics are reported in the supplementary information.