Social learning disfavours the cultural evolution of altruistic punishment

Max Burton-Chellew (maxwell.burton@unil.ch)
University of Lausanne https://orcid.org/0000-0002-5330-0379

Victoire D'Amico
University of Lausanne

Article

Keywords: Cultural evolution, Cultural group selection, Human Cooperation, Institutional selection, Social learning, Social preferences

Posted Date: March 29th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1450110/v1

Social learning disfavours the cultural evolution of altruistic punishment

Maxwell N. Burton-Chellew *[1-2] and Victoire D'Amico [2]

[1] Department of Economics, HEC-University of Lausanne, 1015 Lausanne, Switzerland

[2] Department of Ecology and Evolution, University of Lausanne, 1015 Lausanne, Switzerland.

*Corresponding author: maxwell.burton@unil.ch

maxburtonchellew.com

Keywords

Cultural evolution; Cultural group selection; Human Cooperation; Institutional selection; Social learning; Social preferences

Version date: Wednesday 23rd March 2022.

ABSTRACT

Could human cooperation be the result of evolution favouring groups where altruists punish non-altruists [1-4]? While punishment can benefit groups through increased cooperation, it often costs more than it is worth, challenging group level explanations [1,5,6]. We therefore tested if altruistic punishment could spread among groups via imitation (social learning) rather than direct competition. Groups observed either how punishment increased cooperation, or how it also decreased profits, in a previous experiment. Individuals then voted to either enable or disable all punishment. We found (1) that punishment was common, but most individuals voted to disable punishment and (2) observing how punishment affected payoffs drastically disfavoured punishment, and consequently cooperation. Our results contradict the hypothesis that humans have evolved to cooperate via cultural group selection favouring altruistic punishment [2,3,7-11].

Results from some laboratory experiments have suggested that many humans cooperate and pay to punish non-cooperators even when they or their relatives cannot benefit [1,3,4]. Such punishment is costly but can favour cooperation within groups. Consequently, it has been hypothesized that human cooperation is the evolutionary result of group selection favouring groups with altruistic punishment [1-3,7,8,10,12-15]. However, the same experiments which show that 'altruistic punishment' can sustain cooperation, also show that altruistic punishment is net costly for groups, challenging group level explanations [1,5,6]. Quite simply, the costs of punishment in such experiments, where individuals pay to reduce the earnings of non-cooperators, tend to outweigh the benefits gained from increased cooperation [1,5,6]. Consequently, genetical group selection would be unlikely to favour altruistic punishment.

One potential solution is if cultural groups compete, not directly, but to be preferentially imitated [7,10,16-18]. If groupmembers learn by observing other groups (social learning), then mechanisms/institutions which are correlated with higher levels of cooperation could be preferentially imitated, leading to a process of cultural group selection [7,10,16,17]. However, in the case of altruistic punishment, this would require assuming individuals only learn by observing how punishment affects cooperation, and not payoffs, otherwise social learning too will disfavour excessively costly punishment [7,19]. Therefore, it is important to know if social learning is more likely to favour the imitation of either the most cooperative, or the most successful, groups.

We experimentally tested how social learning affected the level of punishment in a public-goods game where individuals could make costly financial contributions to benefit the group. We enabled social learning by showing individuals data comparing group outcomes from the seminal paper of Fehr & Gächter on altruistic punishment, which included games with and without punishment being enabled [1]. We tested if altruistic punishment could spread when individuals observed how enabling punishment affected average levels of contributions

('cooperation'), and if this spread depended upon whether individuals could also observe how punishment affected average payoffs (specifically, profits).

In our experiment, each session had 16-24 participants (N = 436 participants across 26 sessions in Switzerland). Our No-info treatment had no social learning and these four sessions served as a baseline that replicated the punishment game from Fehr & Gächter [1]. Individuals were repeatedly placed into groups of four for a single round of decision making involving both cooperation and punishment. Individuals were guaranteed to be grouped each time with three people they had not previously interacted with. For each round, individuals first received 20 monetary units (MU, 20 MU = 0.5 Swiss francs) and had to decide how much they wanted to contribute to a group fund that paid out benefits to all group members equally regardless of contribution. Contributions always resulted in a personal loss of 60% but an overall group gain of 60%. Individuals then could pay to induce fines upon specific groupmates, at a cost to fine ratio of 1:3.

Our social learning treatments provided individuals with information about behaviours in the previous experiment of Fehr & Gächter [1]. In each of the five rounds, after the contribution stage and before the punishment stage, we showed individuals either how groups that could punish had higher levels of contributions (Group behaviours treatment, N = six sessions) or we also showed them that despite the higher contributions, groups that could punish also had lower average profits (Group profits treatment, N = six sessions) (Table 1; Extended Data Table 1 & Extended Data Figure 1).

Table 1: A schematic of the social information we showed to participants; see Extended Data Table 1 for precise details.

	Cooperation[1]	Profits[2]
Punishment enabled	Higher	Lower
Punishment disabled	Lower	Higher

[1] Shown in all treatments except the baseline No-info treatment

[2] Only shown in the Group profits and Vote Profits treatments

We found that the number of punishers was significantly lower when individuals could observe that punishment reduced group profits (Extended Data Figure 2, Supplementary Results). Specifically, whereas 87% of individuals punished at least once across the five rounds in each of the No-info and the Group Behaviours treatments, only 74% of individuals punished in the Group Profits treatment (No-info = 59/68, 95% Confidence Interval = [76.5%, 93%] ; Group Behaviours = 87/100, 95% CI = [79%, 92%]; Group Profits = 71/96, 95% CI = [64%, 82%]; Fisher's exact tests comparing between: Group Profits and Group Behaviours, P = 0.029; between Group Profits and No-info, P = 0.052; between Group Behaviours and No-info, P = 1.000).

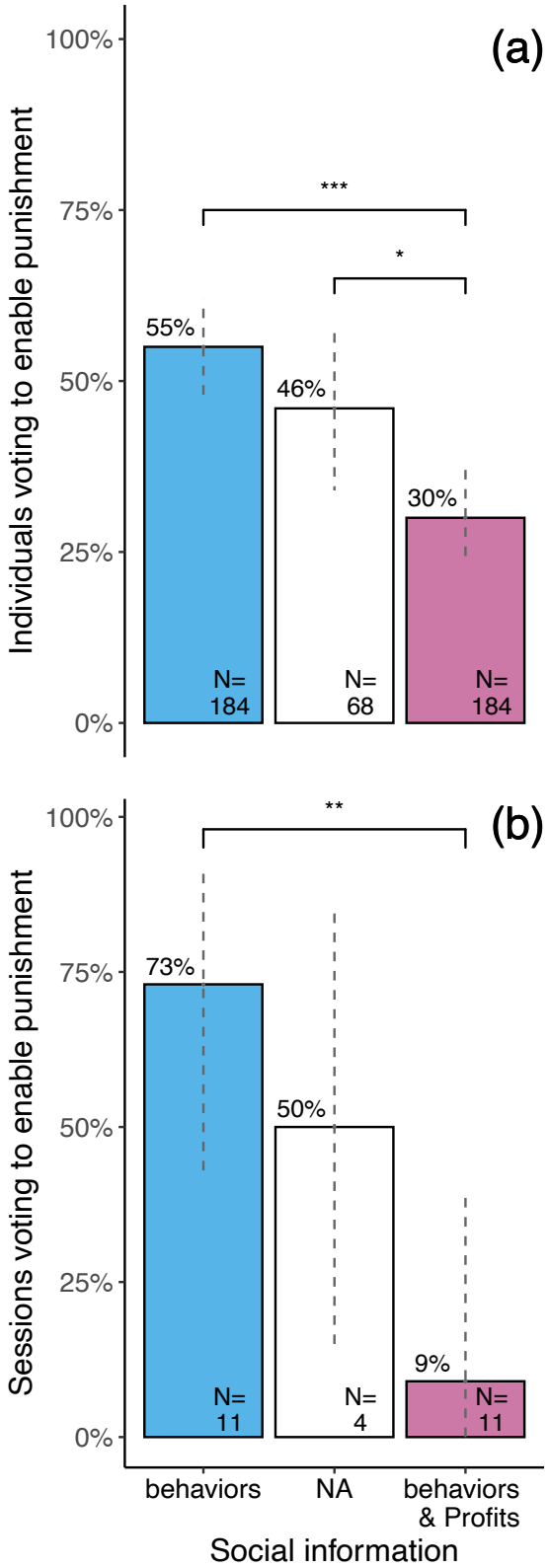
The punishment mechanism typically used in experiments can be thought of as an institution that is adopted by groups or societies to regulate and enforce cooperation [20]. In this case it is important to know if such mechanisms/institutions for altruistic punishment will spread between groups and be maintained. To test how social learning could affect the incorporation and maintenance of such an institution, we made our participants vote to either enable or disable the punishment mechanism. In every session, regardless of treatment, we made individuals vote after the first five rounds of decision making to either enable or disable

punishment for a subsequent and final five rounds of decision making. Punishment was disabled unless more than 50% of individuals voted to enable it. In some sessions, we controlled for game experience by showing all five rounds of the social information and having the vote before the very first round of decision making (N = 5 sessions shown Group Behaviours [Vote Behaviours treatment] and 5 sessions shown Group Profits [(Vote Profits treatment]).

We found no clear support for punishment in our baseline treatment (No-info). Even though 87% of individuals punished at least once, only 46%, significantly fewer, voted in favour of keeping punishment enabled for the second game series (N=31/68, 95% CI = [34%, 57%]; Fisher's exact test on 59/68 versus 31/68, $P < 0.001$, Figure 1a). Consequently, 50% of the No-info sessions decided to disable punishment (Figure 1b; N = 2/4, 95% CI = [15%, 85%]), with total votes for punishment per session ranging from 25% to 69% (from 4/16 to 11/16 individuals).

We next tested how social information on group comparisons affected voting preferences for punishment. Here we include the extra ten sessions that showed all the social information entirely before round one, rather than in-between each round. We found that the effect of social information upon the first round of voting depended on the level of social information available. Whereas observing social information on group behaviours had no significant effect, observing social information on group profits significantly reduced votes for punishment. Specifically, when individuals could only observe how punishment affected levels of contributions (Group Behaviours and Vote Behaviours treatments), 55% voted in favour of punishment (Figure 1a; N=101/184, 95% CI = [48%, 62%]), which was not significantly more votes in favour than in the No-info treatment (Fisher's exact test, 31/68 versus 101/184, P = 0.203). Consequently, 73% of sessions voted to enable punishment (Figure 1b; N = 8/11, 95% CI = [43%, 91%]), although the outcomes were typically close, ranging from 44%-69% in favour (N = 7/16 to 11/16).

In contrast, when individuals could also observe how punishment affected average profits only 30% voted to enable punishment ((Group Profits and Vote Profits treatments combined, Figure 1a; N=55/184, 95% CI = [24%, 37%]). This was significantly less than both those that only saw social information on behaviours (Group Behaviours and Vote Behaviours combined, Fisher's exact test, 55/184 versus 101/184, $P < 0.001$) and those in the No-info treatment (Fisher's exact test, 55/184 versus 31/68, $P = 0.025$). Consequently, only 9% of the sessions with social information on profits voted in favour of punishment (Figure 1b; N = 1/11, 95% CI = [<0.01%, 40%]; votes ranged from 2/16 to 10/16, from 12%-62%), which was significantly fewer sessions than with social information on behaviours (Fisher's exact test, 1/11 versus 8/11, $P = 0.008$).



140

141 ***Figure 1a,b. Social learning from group profits reduce support for punishment.*** *Results of the*

142 *first voting round, (a) by individuals and (b) by sessions depending upon the level of social*

information we provided. Only significant pair-wise comparisons are shown. Dashed lines show 95% confidence intervals.

The near total loss of the punishment mechanism/institution across sessions that could observe how punishment affected average profits led to a dramatic collapse of contributions by the end of the experiment (Figure 2). The final contribution level was only 18% (3.5 MU), 95% CI = [8.9%, 32%], significantly less than the 47% (9.3 MU), 95% CI = [28%, 65%], contributions in the treatments shown just how punishment affected contributions (Welch's two sample t-test on mean contributions per session = 2.4, df = 16.0, P = 0.027).

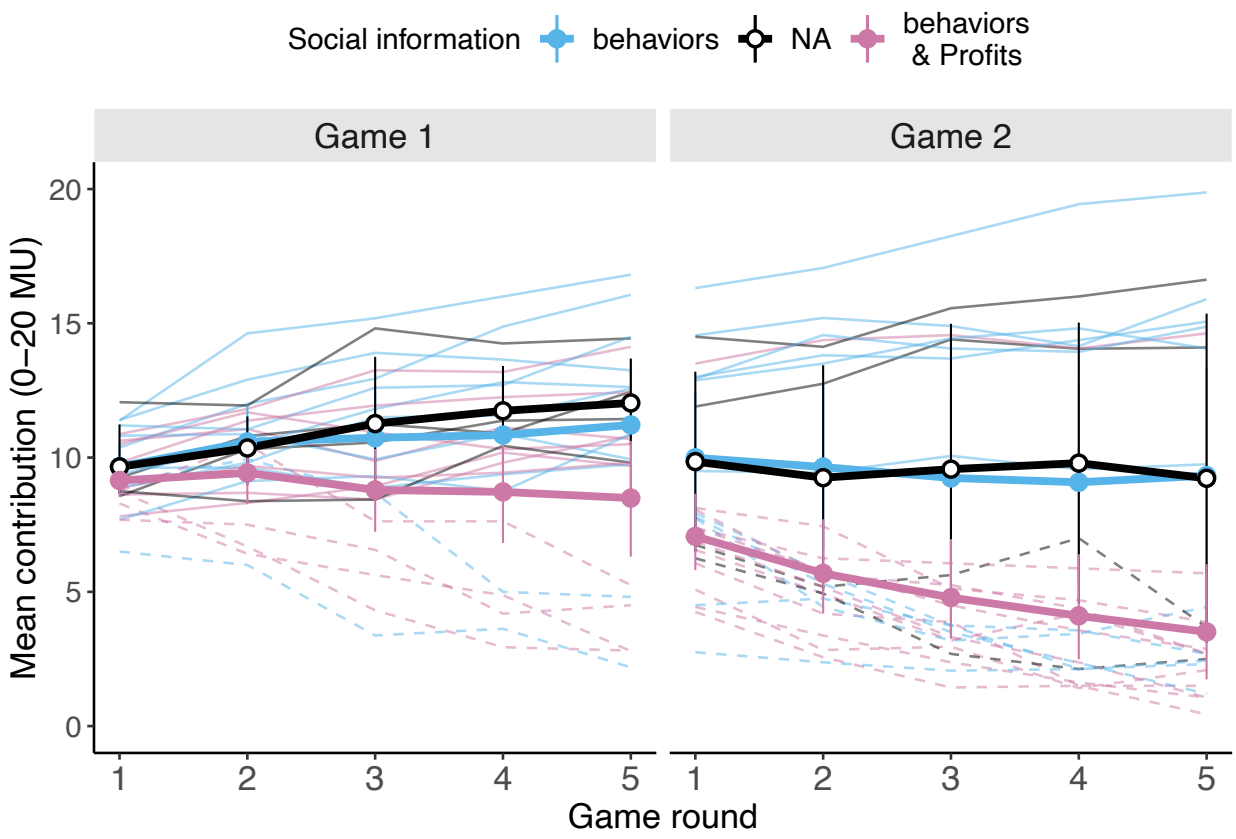


Figure 2. Cultural evolution of contributions to a public good. *Punishment sustained contributions but the loss of punishment led to the loss of contributions. Thick lines connect*

overall mean contributions and 95% confidence intervals. Thin lines show mean contributions per session. Dashed lines mean that punishment had been disabled by a vote.

In summary, we found that cultural evolution via social learning does not favour altruistic punishment and therefore does not favour cooperation. Our central result, that information about group profits led to significantly fewer votes for punishment, did not significantly depend upon whether the vote came before or after game experience (Supplementary Results). We had 26 sessions. Among the 16 sessions that had no initial vote and thus no choice but to initially play with punishment, only seven voted to keep punishment enabled (44%, 95% CI = [23%, 67%]), and none of these seven had been shown social information on profits (two from four in the No-info treatment and five from six in the Group Behaviours treatment, Extended Data Figure 1). Among the 10 sessions that got to vote twice, only two voted to enable punishment in the second game series (20%, 95% CI = [5%, 52%]), one from five in each treatment (Vote Behaviours and Vote Profits, Extended Data Figure 1).

Overall, 20 sessions first played with punishment enabled, and of these 20, only nine voted to keep punishment enabled, an attrition rate of 55%. Of the six sessions that first voted to play without punishment, none of them changed their mind, as all six decided to continue without punishment enabled (Figure 3). Therefore, the punishment 'institution' was lost 55% of the time, but never gained, and in total, only 35% of sessions (N = 9/26, CI = [19%, 54%]), voted to finish the experiment with punishment enabled (Figure 3). Finally, when sessions voted to disable punishment they made more money, further reinforcing the finding that punishment is prohibitively costly (Supplementary Results).

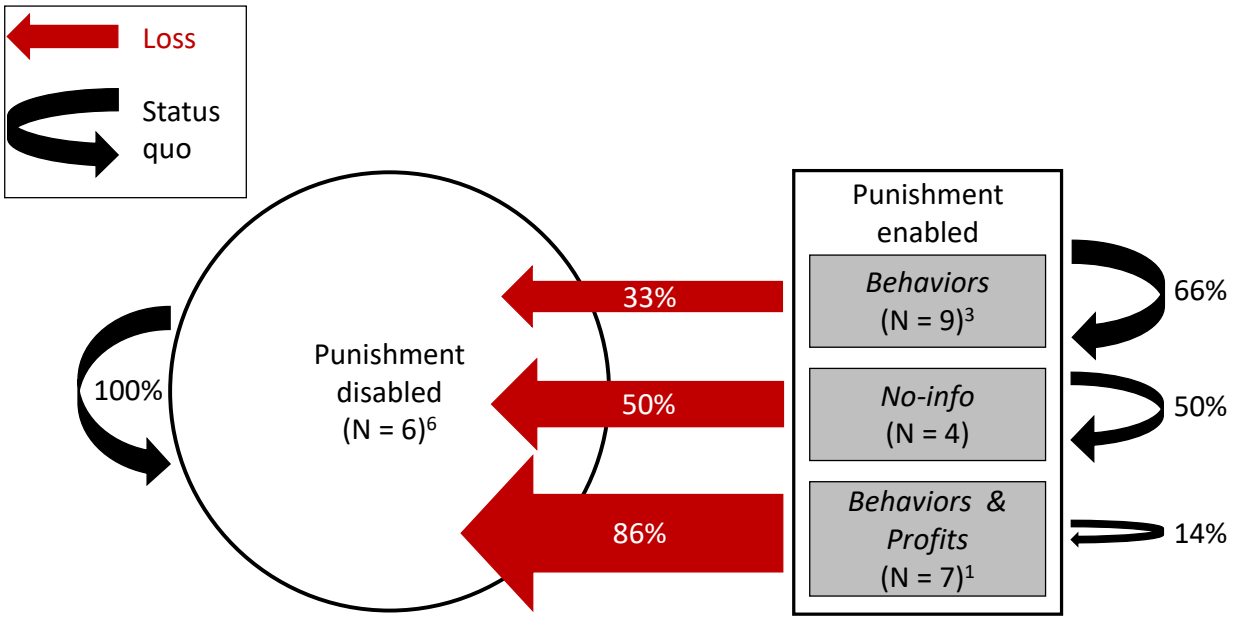


Figure 3. The cultural instability and loss of punishment. *N indicates the number of sessions that started the experiment in that condition. There were 26 sessions in total. Superscript shows the number of sessions that had a vote prior to the first game series and chose to play either with or without punishment enabled. Arrows pointing left (red) indicate a vote to disable punishment and curved arrows (black) indicate a vote for status quo (no change). No sessions that played with punishment disabled subsequently voted to enable punishment.*

In conclusion, our results contradict the hypothesis that social learning could favour the evolution of altruistic punishment or that humans prefer institutions for altruistic punishment [18,21]. For example, even though most individuals punished in the No-info treatment, we found most also voted to disable punishment after experience of the game. Consequently, institutions for altruistic punishment will struggle to have a competitive advantage [21]. Altruistic punishment is not only costly for the individual punishers, but also for the group, and even still for those groups that benefit from higher cooperation, therefore payoff-based learning will impede the spread of altruistic punishment. One possibility is that groups are not democratic/egalitarian like in our experiment, and instead powerful or persuasive individuals are able to implement

punishment institutions [22]. Whether this would be to the benefit of the whole group, rather than just the elites, is unclear [22,23]. Of course, another possibility is that natural punishment is not altruistic [6,24-33].

ONLINE METHODS

Data availability statement

All the data and analysis files are available on the Open Science Framework [34]. A full copy of the experimental instructions translated into English are included here at the end in Appendix 1.

Location, participants, and software

We used z-Tree to program and run the experiment [35]. We conducted all 26 sessions at the University of Lausanne (UNIL) HEC-LABEX facility. LABEX recruited the participants using the ORSEE software and excluded participants that had previously experienced any experiments with punishment [36]. Participants were told that a level of B2 French was necessary. In total we had 436 participants (Female = 251; Male = 175; Other = 3; Declined to respond = 7) who were typically aged 18-25 (Under 21 = 261; 21-25 = 147; 26-30 = 13; 31-35 = 4; Over 35 = 7; Declined to respond = 4).

The public goods game

We replicated the public goods game of Fehr & Gächter [1]. Individuals were put into groups of four each round. We gave individuals 20 Monetary Units (MU, 1 MU = 0.025 Swiss Francs) that they could contribute any amount of to a group fund. All contributions were multiplied by 1.6 before being divided equally among the four group members regardless of differences in contributions. This meant each MU contributed returned 0.4 MU per group members and was therefore personally costly, necessitating a personal loss of 0.6 MU while benefiting the other group members by 1.2 MU in total, leading to a group profit of 0.6 MU. The individual income maximizing decision in the one-shot game is therefore to contribute 0 MU, and the group income maximizing decision is to contribute fully (20 MU).

Each session had 16-24 participants and the game was played in two series of five rounds. Individuals were anonymous and knew that they could never interact with the same person more than once across the 5 rounds of the game ('perfect stranger' matching mechanism). Sessions therefore resembled 'societies' of 16-24 individuals that formed transient mini-groups of four for ephemeral cooperative interactions. When participants played the first game series, they did not know there would be a second game series. When they played the second game series, they had been told that this was the final part of the experiment (which also coincided with the expected duration of the experiment).

Punishment

We used the same punishment mechanism as Fehr & Gächter [1]. After individuals made their contributions, we showed them feedback detailing the individual decisions and incomes of all four groupmembers. Individuals could then pay to make deductions to other members incomes (to 'punish'). For every 1 MU they paid, they deducted 3 MU from another member's account. This meant that for every unit of punishment spending, the group lost 4 MU. Individuals' could punish any and all of their groupmembers, but were limited to a maximum spending of 10 MU per targeted individual. To prevent differences in income from the public-good game affecting individual's ability to afford punishment, we endowed all individuals with 30 MU in the punishment phase. This meant the punisher always had to realize the costs of the punishment they inflicted upon others. However, victims of punishment could only lose the MU they had earned from the public good game stage, which could range from 8-44 MU. Received deductions/punishment could not lead to negative earnings.

Social information on group comparisons

All individuals received feedback on the contributions and payoffs of their own group members. To facilitate social learning about how punishment affects cooperation and relative payoffs, and to mimic groups/societies learning from other societies, we also provided social information comparing outcomes for groups that played with or without punishment in the experiment of

Fehr & Gächter [1]. Their experiment had 10 sessions of 20-24 participants that played two series of public goods games, either with or without punishment for either five or six rounds. We calculated three variables for each of their first five rounds: the mean average per capita (1) contribution, (2) spending on punishment, and (3) profit, which was the average payoff, including gains from contributions and losses from punishment, relative to starting endowment, calculated from averaging across all 10 sessions depending on whether they could punish or not (Extended Data Table 1). This social information consistently showed that groups with punishment contributed more but also had lower profits. While we took the overall average, the relationship between punishment and contributions, and between punishment and profits, was qualitatively the same in each of all 10 sessions of Fehr & Gächter [1] (Extended Data Table 2).

Voting

All sessions played two series of public good games, both for five rounds, either with or without punishment, and either with or without social information (Figure 1). After the first game series, and just before the second game series, all individuals within a session (N = 16-24) had to vote, anonymously. Their decision was to vote on whether to enable punishment or not in the subsequent game series. Punishment was only enabled if more than 50% of that session's participants voted for punishment. After the voting had finished, we informed individuals of the outcome of the vote (punishment enabled or not), but we did not provide information on how close the vote was. This was to prevent, among sessions with the same outcome, variation in vote counts from affecting beliefs about groupmates. In some sessions, there was also a vote before the first game series (Extended Data Figure 1). This was to control for the effect of game experience and learning about the behaviour of other participants in the session before voting.

Experimental procedure and treatments

Regardless of treatment, all participants begun with on-screen instructions detailing the public good game. We based our instructions and control questions on [37]. Participants then read about the punishment mechanism, with instructions and control questions adapted from [4].

Then the procedure differed depending on treatment. In total we had five treatments (Extended Data Figure 1). All participants within a session experienced the same treatment.

Treatment 1: No-info

Our No-info treatment had no social information on group comparisons from the previous experiment and thus served as a baseline measure for comparisons. Participants first played five rounds of the public goods game with punishment. In each round individuals received information on the contributions and earnings of each of their groupmates before deciding if they wanted to punish. After five rounds they then voted to either enable or disable punishment for a second series of five rounds. Our No-info treatment allowed us to measure rates of punishment and how willing groups/societies are to maintain an altruistic punishment mechanism/institution once they have experienced it.

Treatment 2: Group Behaviours

This was the same as our No-info treatment except that we also showed individuals social information comparing group behaviours in the previous experiment of Fehr & Gächter [1]. Specifically, we showed participants the mean average contribution and punishment. We timed this in-between the contribution and the punishment stages of each round of the first game series, using the corresponding round from Fehr & Gächter [1]. This social information had the feature that it always showed, regardless of round, that groups that could punish had higher mean contributions (Extended Data Table 1). After five rounds of game play with social information, the participants voted to either enable or disable punishment for a second and final game series, knowing that there would be no more social information from the previous experiment.

Treatment 3: Group Profits

This was the same as the Group Behaviours treatment except that the social information now contained an extra column, detailing the average profit depending upon whether punishment was enabled or not. This social information had the feature that it always showed, regardless of

round, that while groups that could punish had higher mean contributions, they also always had lower mean profits, sometimes even negative profits (Extended Data Table 1). Again, just like in our No-info and Group Behaviour treatments, the participants voted after round five to either enable or disable punishment for a second game series, knowing that there would be no more social information from the previous experiment.

Treatments 4 & 5: Vote Behaviours & Vote Profits

We also had two control treatments where the first round of voting could not be influenced by experience of the game or groupmates. Instead, we showed participants all five rounds of the relevant social information (either on group behaviours or group profits) before the game and then made them vote. They then played for five rounds with punishment either enabled or disabled accordingly, but with no more social information from the previous experiment, before voting again for a second time for a second game series.

Financial incentives

Individuals received 10 CHF for showing up to the experiment. Then their decisions were financially incentivized. Each MU was worth 0.025 CHF, meaning 20 MU was worth 0.5 CHF. In total, we gave individuals 50 MU in each round (20 MU for the contribution stage and 30 MU for the punishment stage) and 500 MU across 10 rounds of decision making. This meant we gave them a total endowment of 12.50 CHF. Final earnings inclusive of show up fee were rounded up to the nearest CHF and ranged from 19 to 31 CHF, with a mean of 23.3 CHF and a median of 23.0 CHF.

Ethics statement

We obtained ethical approval from the HEC-LABEX ethics committee, who forbid deception in experiments. Before starting each session, all participants signed a consent form. Participants were told that they could leave at any time but that this could forfeit their earnings from the experiment.

Statistical analyses

We used R Studio for analyses [38]. All significance tests are two tailed.

References

1 Fehr, E. & Gachter, S. Altruistic punishment in humans. *Nature* **415**, 137-140 (2002).

2 Fehr, E. & Fischbacher, U. The nature of human altruism. *Nature* **425**, 785-791 (2003).

3 Henrich, J. *et al.* Costly punishment across human societies. *Science* **312**, 1767-1770, doi:10.1126/science.1127333 (2006).

4 Herrmann, B., Thoni, C. & Gachter, S. Antisocial punishment across societies. *Science* **319**, 1362-1367, doi:Doi 10.1126/Science.1153808 (2008).

5 Egas, M. & Riedl, A. The economics of altruistic punishment and the maintenance of cooperation. *Proceedings of the Royal Society B-Biological Sciences* **275**, 871-878, doi:Doi 10.1098/Rspb.2007.1558 (2008).

6 Burton-Chellew, M. N. & Guérin, C. Decoupling cooperation and punishment in humans shows that punishment is not an altruistic trait. *Proceedings of the Royal Society B: Biological Sciences* **288**, 20211611, doi:doi:10.1098/rspb.2021.1611 (2021).

7 Henrich, J. & Boyd, R. Why people punish defectors - Weak conformist transmission can stabilize costly enforcement of norms in cooperative dilemmas. *J Theor Biol* **208**, 79-89, doi:DOI 10.1006/jtbi.2000.2202 (2001).

8 Boyd, R., Gintis, H., Bowles, S. & Richerson, P. J. The evolution of altruistic punishment. *P Natl Acad Sci USA* **100**, 3531-3535 (2003).

9 Zefferman, M. R. & Mathew, S. An evolutionary theory of large-scale human warfare: Group-structured cultural selection. *Evolutionary Anthropology: Issues, News, and Reviews* **24**, 50-61, doi:https://doi.org/10.1002/evan.21439 (2015).

10 Richerson, P. *et al.* Cultural group selection plays an essential role in explaining human cooperation: A sketch of the evidence. *Behav Brain Sci* **39** (2016).

11 Henrich, J. & Muthukrishna, M. The Origins and Psychology of Human Cooperation. *Annual Review of Psychology, Vol 72* **72**, 207-240 (2021).

12 Gintis, H., Bowles, S., Boyd, R. & Fehr, E. Explaining altruistic behavior in humans. *Evol Hum Behav* **24**, 153-172 (2003).

13 Bowles, S. & Gintis, H. The evolution of strong reciprocity: cooperation in heterogeneous populations. *Theor Popul Biol* **65**, 17-28 (2004).

14 de Quervain, D. J. F. *et al.* The neural basis of altruistic punishment. *Science* **305**, 1254-1258 (2004).

15 Fowler, J. H. Altruistic punishment and the origin of cooperation. *P Natl Acad Sci USA* **102**, 7047-7049 (2005).

16 Boyd, R. & Richerson, P. J. Group beneficial norms can spread rapidly in a structured population. *J Theor Biol* **215**, 287-296, doi:10.1006/jtbi.2001.2515 (2002).

17 Henrich, J. Cultural group selection, coevolutionary processes and large-scale cooperation. *J Econ Behav Organ* **53**, 3-35, doi:Doi 10.1016/S0167-2681(03)00094-5 (2004).

18 Henrich, J. Cooperation, punishment, and the evolution of human institutions. *Science* **312**, 60-61, doi:10.1126/science.1126398 (2006).

19 Smith, D. Cultural group selection and human cooperation: a conceptual and empirical review. *Evolutionary Human Sciences* **2**, e2, doi:10.1017/ehs.2020.2 (2020).

20 Richerson, P. J. & Henrich, J. Tribal Social Instincts and the Cultural Evolution of Institutions to Solve Collective Action Problems. *Cliodynamics* **3**, 38-80 (2012).

21 Gurerk, O., Irlenbusch, B. & Rockenbach, B. The Competitive Advantage of Sanctioning Institutions. 10.1126/science.1123633. *Science* **312**, 108-111 (2006).

22 Watts, J., Sheehan, O., Atkinson, Q. D., Bulbulia, J. & Gray, R. D. Ritual human sacrifice promoted and sustained the evolution of stratified societies. *Nature* **532**, 228-231, doi:10.1038/nature17159 (2016).

23 Henrich, J., Chudek, M. & Boyd, R. The Big Man Mechanism: how prestige fosters cooperation and creates prosocial leaders. *Philos T R Soc B* **370** (2015).

24 Gardner, A. & West, S. A. Cooperation and punishment, especially in humans. *Am Nat* **164**, 753-764 (2004).

25 Barclay, P. Reputational benefits for altruistic punishment. *Evol Hum Behav* **27**, 325-344, doi:Doi 10.1016/J.Evolhumbehav.2006.01.003 (2006).

26 Dawes, C. T., Fowler, J. H., Johnson, T., McElreath, R. & Smirnov, O. Egalitarian motives in humans. *Nature* **446**, 794-796 (2007).

27 Pedersen, E. J., Kurzban, R. & McCullough, M. E. Do humans really punish altruistically? A closer look. *Proceedings of the Royal Society B-Biological Sciences* **280**, doi:10.1098/Rspb.2012.2723 (2013).

28 Kurzban, R. & DeScioli, P. Adaptationist punishment in humans. *Journal of Bioeconomics* **15**, 269-279, doi:10.1007/s10818-013-9153-9 (2013).

29 Raihani, N. J. & Bshary, R. The reputation of punishers. *Trends in Ecology & Evolution* **30**, 98-103, doi:10.1016/j.tree.2014.12.003 (2015).

30 Hoeft, L. & Mill, W. Selfish punishers: An experimental investigation of designated punishment behavior in public goods. *Econ Lett* **157**, 41-44 (2017).

31 Pedersen, E. J., McAuliffe, W. H. B. & McCullough, M. E. The Unresponsive Avenger: More Evidence That Disinterested Third Parties Do Not Punish Altruistically. *J Exp Psychol Gen* **147**, 514-544, doi:10.1037/xge0000410 (2018).

32 Molho, C., Tybur, J. M., Van Lange, P. A. M. & Balliet, D. Direct and indirect punishment of norm violations in daily life. *Nat Commun* **11**, doi:ARTN 3432 10.1038/s41467-020-17286-2 (2020).

33 Deutchman, P., Bracic, M., Raihani, N. & McAuliffe, K. Punishment is strongly motivated by revenge and weakly motivated by inequity aversion. *Evol Hum Behav* **42**, 12-20, doi:10.1016/j.evolhumbehav.2020.06.001 (2021).

34 Burton-Chellew, M. & D'Amico, V. Data and files for: Social learning disfavors the cultural evolution of altrustic punishment. *Open Science Framework: https://osf.io/h6cb3/*, doi:https://osf.io/h6cb3/ (2022).

35 Fischbacher, U. z-Tree: Zurich toolbox for ready-made economic experiments. *Exp Econ* **10**, 171-178, doi:Doi 10.1007/S10683-006-9159-4 (2007).
36 Greiner, B. Subject pool recruitment procedures: organizing experiments with ORSEE. *J Econ Sci Assoc* **1**, 114-125, doi:10.1007/s40881-015-0004-4 (2015).
37 Fischbacher, U. & Gachter, S. Social Preferences, Beliefs, and the Dynamics of Free Riding in Public Goods Experiments. *Am Econ Rev* **100**, 541-556, doi:10.1257/aer.100.1.541 (2010).
38 Team, R. *RStudio: Integrated Development Environment for R*. (http://www.rstudio.com, 2020).

Supplementary Information for:

Social leaning disfavours the cultural evolution of altruistic punishment

Maxwell N. Burton-Chellew *[1-2] and Victoire D'Amico [2]

[1] Department of Economics, HEC-University of Lausanne, 1015 Lausanne, Switzerland

[2] Department of Ecology and Evolution, University of Lausanne, 1015 Lausanne, Switzerland.

*Corresponding author: maxwell.burton@unil.ch

maxburtonchellew.com

Version date: Monday 14th March 2022

SUPPLEMENTARY RESULTS

Information on payoffs reduced punishment

As reported in the main text, we found that the number of punishers was significantly lower when individuals could observe that punishment reduced group profits (Extended Data Figure 2). The mean spending on punishment was also significantly lower in the Group Profits treatment. Specifically, it was 2.7 MU 95% bootstrapped CI = [2.4, 3.1,] in the No-info treatment, 2.3 MU 95% bootstrapped CI = [2.1, 2.6] MU in the Group Behaviours treatment and 2.2 MU 95% bootstrapped CI = [1.8, 2.5]] in the Group Profits treatment (linear mixed model controlling for participant, $F_{2,264} = 9.0$, $P < 0.001$; pairwise comparisons between: Group Profits and No-info, t=3.7, $P < 0.001$; Group Profits and Group Behaviours, t=-3.6, $P < 0.001$; Group Behaviours and No-info, t=0.4, P = 0.674).

Because social information reduces the level of punishment, then this may in turn reduce the level of cooperation, and this in turn may increase the levels of punishment, we also examine

these effects in the opening round, before any effects upon levels of punishment can affect subsequent behaviour. There were significant differences in the opening round between the Group Profits and the Group Behaviours treatments. Specifically, only 44% of individuals chose to punish in the first round of the Group Profits treatment (N = 42/96, 95% CI = [34%, 54%]), with a mean spending on punishment of 1.6 MU, 95% bootstrapped CI = [1.0, 2.3]; compared to 51% choosing to punish and spending 2.2 MU, 95% bootstrapped CI = [1.5, 3.0] in the No-info treatment (N = 35/68, 95% CI = [40%, 63%]), and 60% choosing to punish and spending 2.1 MU, 95% bootstrapped CI = [1.6, 2.7] in the Group Behaviours treatment (N = 60/100, 95% CI = [50%, 69%]) (statistics comparing Group Profits versus Group Behaviours treatments: Fisher's Exact Test comparing frequency of punishers, P = 0.0317; Linear model comparing levels of spending, t = 3.6, P < 0.0001. Comparisons with the No-info treatment were non-significant).

Game experience does not increase support for punishment

Our central result, that information about group profits led to significantly fewer votes for punishment, did not significantly depend upon whether the vote came before or after game experience. Comparing among participants, we find a significantly negative effect of social information containing profits both before game experience (Fisher's exact test, comparing 30% in the Vote Profits treatment to 52% in the Vote Behaviours treatment, 26/88 versus 44/84 votes, P = 0.003) and after game experience (Fisher's exact test, comparing 30% in the Group Profits treatment versus 57% in the Group Behaviours treatment, 29/96 versus 57/100 votes, P < 0.001).

Participants in the Vote Behaviours and Vote Profits treatments got to vote two times, once before each game series, which allowed us to test within participants. Around 20% of participants changed their vote, specifically 21% changed in the Vote Behaviours treatment (N = 18/84) and 18% in the Vote Payoffs treatment (N = 16/88, Fisher's exact test, P = 0.702). Again, we found that experience did not matter, as individual support for punishment did not significantly change after game experience in either treatment (Vote Behaviours: support went

from 52% to 45%, from 44 to 38/84 votes, Fisher's exact test, P = 0.440; Vote Profits: support went from 30% to 36%, from 26 to 32/88 votes, Fisher's exact test, P = 0.423).

Punishment was costly

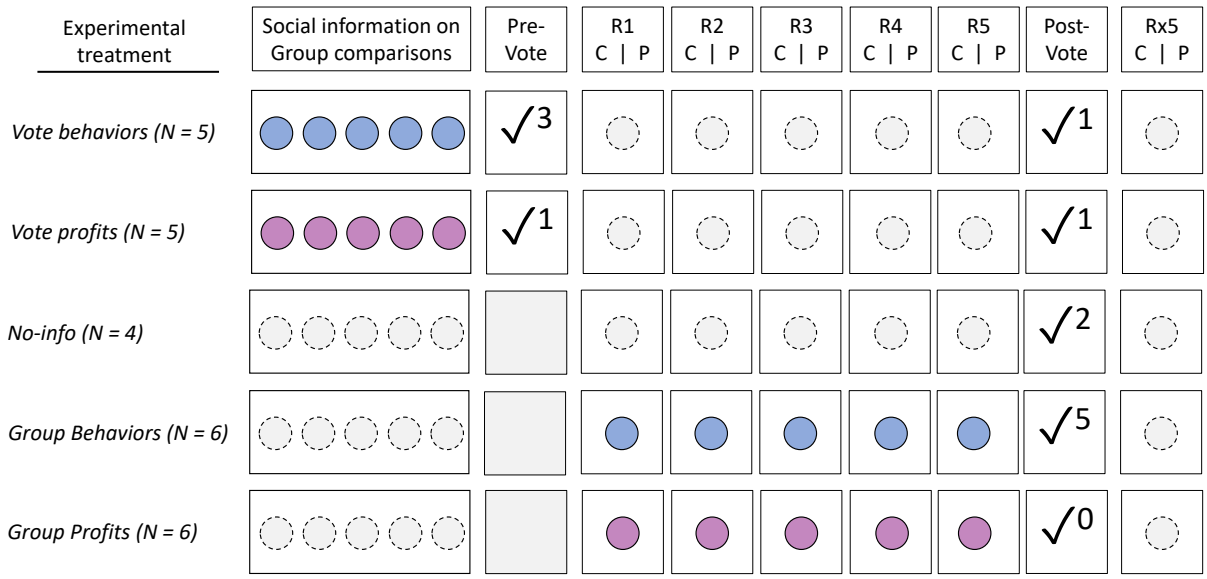
Specifically, when sessions played without punishment (N = 23), they made an average profit per capita per round of 2.8 MU, 95% CI = [2.43, 3.16], whereas when sessions played with punishment enabled (N = 29), they not only earned significantly less, but they made a net loss on average per capita per round of -2.2 MU, 95% CI = [-3.41, -1.06] (Welch two sample t-test, t = 7.7, df = 33.2, P < 0.001).

Extended Data Table 1: *Social information comparing groups with and without punishment. We showed individuals these data, one round at a time, from Fehr & Gächter [1] experiment. The mean profit was only shown in some treatments.*

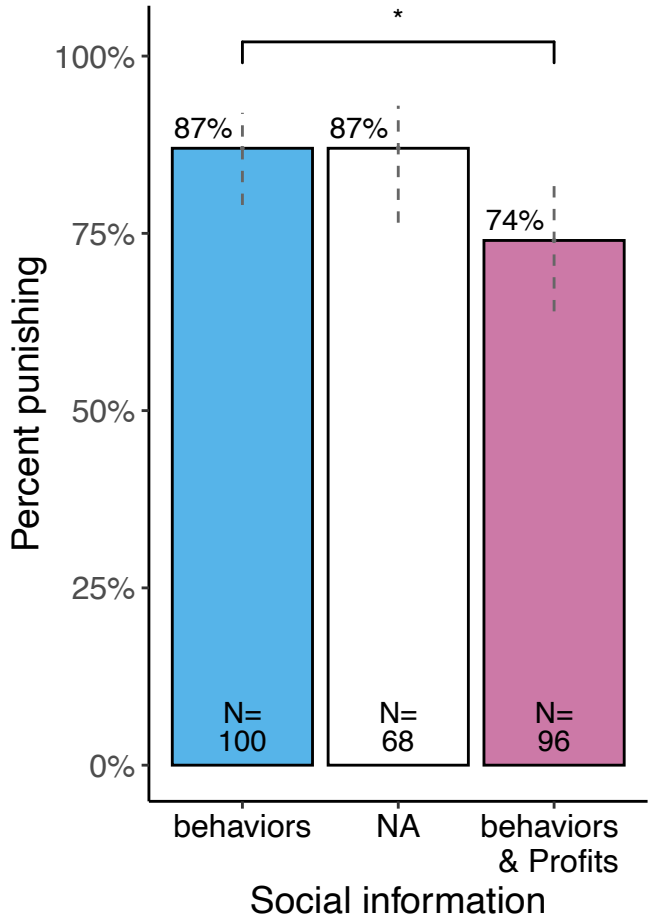
Round	Punishment enabled	Mean contribution	Mean punishment	Mean Profit
Round 1	No	10.0	0	6.0
	Yes	12.6	2.1	-0.5
Round 2	No	8.8	0	5.3
	Yes	13.9	2.4	-1.2
Round 3	No	7.4	0	4.4
	Yes	14.8	2.1	0.5
Round 4	No	6.2	0	3.7
	Yes	15.2	1.9	1.8
Round 5	No	5.6	0	3.4
	Yes	15.8	1.9	2.1

Extended Data Table 2: *Data by session from the previous experiment of Fehr & Gächter* [1].

Session	Punishment Institution	Mean contribution (0-20 MU)	Mean punishment	Mean Profit
1	No	7.5	0	4.5
1	yes	16.2	2.5	-0.2
2	No	5.25	0	3.15
2	yes	13.3	1.4	2.3
3	No	3.6	0	2.2
3	yes	14.3	1.7	2.0
4	No	7.3	0	4.4
4	yes	15.9	2.0	1.6
5	No	7.7	0	4.6
5	yes	14.4	2.3	-0.1
6	No	8.2	0	4.9
6	yes	12.4	2.6	-2.6
7	No	7.2	0	4.3
7	yes	15.4	2.1	1.0
8	No	8.9	0	5.3
8	yes	16.4	1.8	2.7
9	No	9.4	0	5.65
9	yes	16.1	2.8	-1.0
10	No	6.1	0	3.6
10	yes	11.6	1.6	0.7



Extended Data Figure 1. Experiment design. *N indicates the number of sessions. R1-5 = game round of each game series. C = Contribution phase, P = Punishment phase. Voting determined if the punishment mechanism was enabled or not. The superscript numeral indicates how many sessions subsequently enabled punishment in the relevant game series. Social information, from Extended Data Table 1, summarized either behaviours only (blue) or behaviours and profits (pink) from each round of a previous experiment. The five rounds of social information were shown either in total before the first game series (top two treatments), or round by round in-between the contribution and punishment phases of each round (bottom two treatments). No social information was ever shown in the second game series.*



Extended Data Figure 2. Observing group profits reduces the number of punishers. *The percentage of individuals that punished somebody at least once in any of the five rounds from the first game series, along with 95% confidence intervals and significant pairwise comparisons. All these sessions had been randomly assigned to play with the punishment mechanism enabled. All individuals saw, in each round, the contributions and payoffs of their groupmates, plus an additional level of social information from the previous experiment of Fehr & Gächter [1] (Social information, Table 1). All non-shown pair-wise comparisons were non-significant. Dashed lines show 95% confidence intervals.*

Appendix 1: full copy of experimental instructions

A. General Instructions (displayed to every participant)

<u>Screen 1, General Instructions</u>

You are going to take part in an economic experiment that has been funded solely for academic research purposes.

If you read the following instructions carefully, you can earn a reasonable amount of money depending on your decisions.

It is therefore very important that you read these instructions carefully. You have all the time you need.

According to the HEC-Labex rules, experimenters are not allowed to lie to you. Therefore, all information given to you during the experiment is true. At no time during the experiment will we lie to you or try to deceive you. For example, if we tell you that you are going to play with a real person, you will really be playing with a real person.

You are not allowed to communicate during the experience. Breaking this rule will exclude you from the experience and will deprive you of any payment. If you have any questions, please don't hesitate to ask us. Please raise your hand. A member of the experiment team will come and answer you privately.

During the experiment, we will not talk about CHF but rather about MUs (Monetary Units).
During the experiment, your earnings will be calculated in MUs.
At the end of the experiment, the total amount you have earned will be converted into CHF with the following rate:

1 MU = 0.025 CHF, so 20 MU = 0.5 CHF.

At the end of the experiment, the generated amount will be immediately paid to you in cash.

We will describe the experimental process on the next page.

<u>Screen 2, Experimental process and instructions for the contribution stage</u>

You will be in a group of 4 people, consisting of yourself and 3 other real people.

There will be 5 rounds of individual decision making. Each decision is anonymous. In each of the 5 rounds, you will be randomly placed in a new group with different people. You will never be grouped with the same person twice during the 5 rounds. There will never be two people in the same group more than once. There are X (*according to the session*) people participating in

the experiment. You will be grouped with 3 different people 5 times, which means that you will interact with 15 different people in total.

For each decision round, there will be 2 steps (step 1 and step 2).

We will first explain Step 1 in full. Then you will have some comprehension questions that will help you understand the instructions (this is not a test!).

We will then explain Step 2 and the experiment will begin.

Step 1: The Contribution Decision

Each member of the group of 4 must decide how to divide 20 MUs. You can keep these 20 MUs in your private account or you can contribute all or part of them to a project. Each MU that you do not put into the project will automatically remain in your private account.

On the next page you can see how your income from step 1 will be calculated.

<u>Screen 3, Instructions for the contribution stage</u>

Step 1: The Contribution Decision

Your Private Account Income:

You will earn 1 MU for every MU you put into your private account.
For example, if you kept 20 MUs in your private account (and therefore did not contribute to the project) your income will be exactly 20 MUs from your private account.
If you keep 6 MU in your account, you will have an income of 6 MU from that account.
Nobody but you receives anything from your private account.

Your project income:

You and your group members will benefit from the amount you put into the project.
In the same way, you will also get an income from the contributions of other group members to the project.
The income of each group member, including you, will be determined as follows:

Project income = sum of all contributions x 0.4

For example, if the sum of all contributions to the project is 60 MU, then you and the other members of your group will receive 60 x 0.4 = 24 MU from the project.
If four group members contribute a total of 10 MU to the project, you and the other members will each receive 10 x 0.4 = 4 MU.

Total income:

Your total income is the sum of your income from your private account and your income from the project :

Income from your private account (= 20 - project contribution)
+ Project income (= 0.4 x sum of all project contributions)
= Your total income

You will start each round, along with the other 3 members of your group, by indicating an individual contribution to the project.

This will be the end of the first step of the round. On the next page, you will find questions to help you understand this step.

Screen 5, Understanding question 1

Question 1: Each member of the group receives 20 MU. No one (including you) contributes to the project.
What is your total income (in MU)?
What is the total income of each other group member?

Screen 6, Understanding question 2

Here is the second question

Question 2: Each member of the group receives 20 MU. You contribute 20 MU to the project. Each of the other three group members also contributes 20 MU to the project.
What is your total income (in MU)?
What is the total income of each other group member?

Screen 7, Understanding question 3

Here is the third question

Question 3: Each group member receives 20 MUs. The other three group members contribute a total of 30 MU to the project.
What is your total income (in MU), if you - in addition to the 30 MU - contribute 0 MU to the project?
What is your total income (in MU), if you - in addition to the 30 MU - contribute 10 MU to the project?
What is your total income (in MU), if you - in addition to the 30 MU - contribute 15 MU to the project?

Screen 8, Understanding question 4

Here is the last question.

Question 4: Each member of the group receives 20 MU. Suppose you contribute 8 MU to the project.
What is your total income (in MU) if the other members of the group together - in addition to you 8 MU - contribute an additional 7 MU to the project?
What is your total income (in MU) if the other members of the group together - in addition to you 8 MU - contribute an additional 12 MU to the project?
What is your total income (in MU) if the other members of the group together - in addition to you 8 MU contribute with 22 additional MU to the project?

Screen 9, Instructions for the punishment stage

Step 2: The Punishment Decision

After you have made your contribution decision, you will see on the screen the details of the decisions of each member of your group.

You will be able to decrease the income of each of the other members of your group by assigning punishment points or you can leave the income unchanged. The other members of your group can also decrease your income if they wish.

Assigning Punishment Points on Income:

You will need to decide if you are awarding punishment points to each of the other 3 members of your group, and if so, how many. In either case, you will need to enter a number for each of them. If you do not want to change the income of a specific member of your group, you will have to enter 0.

Each punishment point assigned costs you 1 MU. For example, if you assign 2 punishment points to one member of your group, it will cost you 2 MU; if you also assign 5 punishment points to another member, it will cost you another 5 MU; and if you assign 0 points to the last member of your group, it will cost you nothing more. In total, in this example, you would have awarded 7 punishment points, so your total punishment costs would be 7 MU (2 + 5 + 0).

At the beginning of step 2, you will receive a new sum of 30 MU.
You can use this money to allocate punishment points to your group members.
The maximum number of punishment points you can award for each round is 30 MU and the maximum punishment you can award to a member of your group is 10 MU.

Screen 10, Instructions for the punishment stage

Step 2: The Punishment Decision

Receiving punishment points

If someone inflicts 0 punishment points on a specific member of their group, it will not affect that member's income. However, if someone inflicts 1 punishment point on a specific member of their group, it will reduce that member's income by 3 MU. If someone inflicts 2 punishment points on a member of their group, that member's income will be reduced by 6 MU, and so on. In short, each punishment point inflicted on another member of his group will reduce the income of that member by 3 MU.

The total income lost in the second step depends on the total number of punishment points suffered. If a person, for example, receives a total of 4 punishment points (inflicted by all other members of his group during this period), his final income will be reduced by 12 MU. If a person receives a total of 5 punishment points, their final income is reduced by 15 MU.

The maximum punishment a person can receive cannot exceed their earnings from the first stage. Therefore, if a person earns 27 MUs in the first stage, he or she could only lose 27 MUs of maximum total punishment, inflicted by the other members of his or her group. Thus, it is impossible to end up with a negative balance.

Similarly, the amount a person can spend on inflicting deductions on others is limited to the received amount of 30 MU. This again ensures that it is impossible to end up with a negative balance.

Your total income from both steps is therefore calculated as follows:

Total income from steps 1 and 2 =
Income from Step 1
- 3 x (sum of punishment points incurred)
+ your second amount received (30 MU)
- total cost of punishment points you have inflicted

After the contribution decision step you, along with the other 3 members of your group, will have to indicate the punishment points you will assign to your group members.

This is the end of step 2. On the next page you will find questions to help you understand this second step of the round.

<u>Screen 11, Understanding questions</u>

Please answer the six comprehension questions. They will help you understand how your income is calculated based on your decisions.

There is a calculator available on the right side of the screen. To access the calculator, click on the image at the bottom right of your screen.

Once you have answered, you will see the correct answers on the next page.

Question 1: Suppose that in step 2, you assign the following punishment points to the other 3 group members: 5, 4 and 0 respectively. What is your total cost of awarding these punishment points?

Question 2: What are your costs if you award a total of 0 points?

Question 3: By how much MU will your income in the first stage change if you receive a total of 0 points of punishment inflicted by the other group members?

Question 4: By how much MU will your income in the first stage change if you receive a total of 4 points of punishment from the other group members?

Question 5: By how much MU will your income in the first stage change if you receive a total of 15 points of punishment from the other group members?

Question 6: What is the value of the second amount that you can use to inflict punishment points?

B. Specific Instructions for Group behaviours and Group profit treatments

<u>Screen 1, Instruction for social information – Group behaviours/Profits</u>

More information

The experiment will now begin

During the experiment, you will all see information from an experiment, similar to yours, that took place in the past at another Swiss university. It was published by Fehr & Gaechter in 2002.

In this other experiment all participants performed the same two steps as you (step 1 and step 2) for at least 5 rounds. They also did the experiment another time, either before or after, with only step 1 (punishing was impossible).

Specifically, we will show you and all other participants the following information (individual averages in the past experiment);

1) What was the average contribution decision (0-20 MU) per round (step 1), depending on whether or not they could punish each other (we will show you the average decision, rounded to one decimal place).

2) What was the average number of punishment points per round that an individual gave to the 3 members of his team (step 2), when they could punish each other (we will show you the average decision, rounded to one decimal place).

3) What was the average profit in MU (positive or negative) (at the end of the round), depending on whether they could, or could not, punish each other (We will show you the average profit made, rounded to one decimal place). [**only shown in Group Profits treatment]**

This information will appear between the time you have to make a contribution decision and the time you can punish your group members.
You will see the information for the round of the past experiment that corresponds to the round you are in (as a reminder you will be doing 5 rounds).

Screen 2, Understanding questions for social information (True or False)

Please answer the four comprehension questions. They will help you understand the information you are about to see.

When you have answered them, you will see the correct answers on the next page.
Question 1: The additional information we are going to show you comes from an experience that took place in the past?

Question 2: The additional information we are going to show you is separated into a situation where punishing was possible and a situation where punishing was impossible?

Question 3: The additional information we are going to show you corresponds to the averages of the participants according to the different rounds of the past experiment?

Question 4: Each person sees additional information from different experiences?

C. Specific Instructions for Vote behaviours and Vote profit treatments

Screen 1, Instruction for the vote

Vote

Before we begin, each participant, including you, will cast a vote that will determine how the rest of the experiment will proceed.

You can vote for either of these two options:

1) Participate in the experiment WITH the opportunity for everyone to punish and be punished. This implies that everyone will participate in 5 rounds where everyone will first receive 20 MU. You will have to choose how to distribute these 20 MU as explained before (step 1). Then, everyone will receive 30 MU and can punish the members of their group and be punished (step 2).

2) Participate in the experiment WITHOUT the possibility for everyone to punish and be punished. This implies that everyone will participate in 5 rounds where everyone will first receive 20 MU. You will have to choose how to distribute these 20 MU as explained above (step 1). Then, everyone will receive 30 MU but no one will be able to punish the members of their group or be punished (step 2 blocked).

Each participant will vote and the majority vote will be applied for the rest of the experiment. In case of a tie, the experiment will continue without the possibility of punishing or being punished.

Screen 2, Understanding questions for the vote (True or False)

Please answer the three comprehension questions. They will help you understand the vote.

When you have answered, you will see the correct answers on the next page.

Question 1: Will your vote alone determine the outcome of the experiment?

Question 2: Will punishing be possible regardless of the outcome of the vote?

Question 3: If there is a tie, will the experiment proceed without the possibility of punishment?

Screen 3, Instruction for social information – Vote behaviours/Profits

Voting, additional information

Before you vote, you will all see information from an experiment, similar to yours, that took place in the past at another Swiss university. It was published by Fehr & Gaechter in 2002.

In this other experiment, participants did the experiment once with the possibility of punishing each other and once without this possibility. Half of the participants started with one of the two situations and the other half with the other. The situations were then reversed.

Specifically, we will show you and all other participants this information (individual averages in the past experiment);

1) What was, per round, the average contribution decision (0-20 MU) (step 1), depending on whether they could, or could not, punish each other (We will show you the average decision, rounded to one decimal place).

2) What was the average number of punishment points per round given by an individual to the 3 members of his group (step 2), when they could punish each other (We will show you the average decision, rounded to one decimal place).

3) What was the average benefit in MU (positive or negative) (at the end of the round), depending on whether they could punish each other (We will show you the average benefit realized, rounded to one decimal place).[**only shown in Vote Profits treatment]**

This information will appear on your screen 5 times in a row, corresponding to 5 rounds of the past experiment, then you will have to vote.

Screen 4, Understanding questions for social information (True or False)

See screen B. 2

D. Instructions for the second game series, Treatments voting for the first time

Screen 1, Instruction for the vote

Second part of the experiment

You will now do the experiment a second and final time.

This second experiment is similar to the one you just did (5 rounds and two steps per round) with one exception: before you start, you must take a vote that will determine how this second experiment will be conducted.

Vote

You can vote for either of these two options:

1) Participate in the experiment WITH the opportunity for everyone to punish and be punished. This implies that everyone will participate in 5 rounds where everyone will first receive 20 MU. You will have to choose how to distribute these 20 MU as explained before (step 1). Then, everyone will receive 30 MU and can punish the members of their group and be punished (step 2).

2) Participate in the experiment WITHOUT the possibility for everyone to punish and be punished. This implies that everyone will participate in 5 rounds where everyone will first receive 20 MU. You will have to choose how to distribute these 20 MU as explained above (step 1). Then, everyone will receive 30 MU but no one will be able to punish the members of their group or be punished (step 2 blocked).

Each participant will vote and the majority vote will be applied for the rest of the experiment. In case of a tie, the experiment will continue without the possibility of punishing or being punished.

Screen 2, Understanding questions for the vote (True or False)

See screen C. 2

E. Instructions for the second game series, Treatments voting for the second time

<u>Screen 1, Instruction for the vote</u>

Second part of the experiment

You will now do the experiment a second and final time.

This second experiment is similar to the one you just performed (in 5 rounds and two steps per round). You will again start with a vote that will determine how this second experiment will be conducted.

This time we will not show you any additional information.

Vote

As a reminder, you can vote for either of these two options:

1) Participate in the experiment WITH the opportunity for everyone to punish and be punished. This implies that everyone will participate in 5 rounds where everyone will first receive 20 MU. You will have to choose how to distribute these 20 MU as explained before (step 1). Then, everyone will receive 30 MU and can punish the members of their group and be punished (step 2).

2) Participate in the experiment WITHOUT the possibility for everyone to punish and be punished. This implies that everyone will participate in 5 rounds where everyone will first receive 20 MU. You will have to choose how to distribute these 20 MU as explained above (step 1). Then, everyone will receive 30 MU but no one will be able to punish the members of their group or be punished (step 2 blocked).

Each participant will vote and the majority vote will be applied for the rest of the experiment. In case of a tie, the experiment will continue without the possibility of punishing or being punished.