Deterrence through punishment can resolve collective risk dilemmas in carbon emission games

Luo-Luo Jiang
School of Information Management and Artificial Intelligence, Zhejiang University of Finance and Economics, Hangzhou, 310018, China

Zhi Chen*
Department of Modern Physics, University of Science and Technology of China, Hefei, 230026, China

Matjaž Perc†
Faculty of Natural Sciences and Mathematics, University of Maribor, Koroška cesta 160, 2000 Maribor, Slovenia
Department of Medical Research, China Medical University Hospital, China Medical University, Taichung 404332, Taiwan
Alma Mater Europaea, Slovenska ulica 17, 2000 Maribor, Slovenia
Complexity Science Hub Vienna, Josefstädterstraße 39, 1080 Vienna, Austria and
Department of Physics, Kyung Hee University, 26 Kyungheedae-ro, Dongdaemun-gu, Seoul, Republic of Korea

Zhen Wang
Center for OPTical IMagery Analysis and Learning (OPTIMAL) and School of Mechanical Engineering, Northwestern Polytechnical University, Xi'an 710072, China

Jürgen Kurths
Potsdam Institute for Climate Impact Research (PIK), 14473 Potsdam, Germany
Department of Physics, Humboldt University, 12489 Berlin, Germany and
Institute for Complex Systems and Mathematical Biology, University of Aberdeen, Aberdeen AB24 3UE, United Kingdom

Yamir Moreno‡
Institute for Biocomputation and Physics of Complex Systems (BIFI), University of Zaragoza, 50009 Zaragoza, Spain
Department of Theoretical Physics, University of Zaragoza, 50009 Zaragoza, Spain and
CENTAI Institute, 10138 Turin, Italy

Collective risk social dilemmas are at the heart of the most pressing global challenges we are facing today, including climate change mitigation and the overuse of natural resources. Previous research has framed this problem as a Public Goods Game (PGG), where a dilemma arises between short-term interests and long-term sustainability. In the PGG, subjects are placed in groups and asked to choose between cooperation and defection, whilst keeping in mind their personal interests as well as the commons. Here we explore how and to what extent the costly punishment of defectors is successful in enforcing cooperation by means of human experiments. We show that an apparent irrational underestimation of the risk of being punished plays an important role, and that for sufficiently high punishment fines this vanishes and the threat of deterrence suffices to preserve the commons. Interestingly, however, we find that high fines not only avert freeriders, but they also demotivate some of the most generous altruists. As a consequence, the tragedy of the commons is predominantly averted due to cooperators that contribute only their 'fair share' to the common pool. We also find that larger groups require larger fines for the deterrence of punishment to have the desired prosocial effect.

With worsening global climate change, how to reduce carbon emissions has become a challenge which is a collective risk dilemma, displaying the contradiction between short-sighted behavior and farsighted behavior of human beings. To avoid the problem of second-order free-riding, we introduced costly punishment with deterrence where all payers in the same group pay for the punishment. In the carbon emission games, we quantitatively studied the role of deterrence in promoting cooperation. When the risk of being punished is not large enough, players tend to obviously underestimate the possibility of punishment, leading to the wide spreading of myopic behavior. This is why people tend to do nothing in carbon emission games, making global warming more and more serious. Only when the probability of being punished is high enough, the effective deterrent arises, where most players become cooperators who only contribute "fair share" to the public pool, thus avoiding the tragedy of public of carbon emissions. This provides important enlightenment for solving the problem of collective risk dilemmas such as carbon emissions.

* chenzyn@ustc.edu.cn
† matjaz.perc@gmail.com
‡ yamir.moreno@gmail.com

I. INTRODUCTION

Social dilemmas are common in human society. For example, it is well-known that human activities have already

changed global climate through the emission of greenhouse gases, specially CO_2, into the atmosphere [1, 2]. If we do not reduce the release of these greenhouse gases, much greater changes, such as global warming and sea-level rise, will become inevitable consequences [3–6]. To this end, intergovernmental cooperation and coordination are necessary [7–10]. A dilemma thus naturally poses itself: a severe reduction might depress economy and lead to less short-term economic benefits, whereas implementation of insufficient -or no- measures might cause severe climate changes and huge economic losses in the mid to the long term.

To mimic this type of social dilemmas, a number of game models stylizing the climate change problem with countries or governments as players, have been proposed during the past years [11–15]. Typical examples include: threshold public goods games, requiring a minimal investment into a common pool [16]; emission games, where each actor can only release a certain amount of CO_2 per year [17]; climate negotiation games, which need a special negotiation scenario [18]; dynamic climate-change games, involving stochasticity and scientific uncertainty [19] and collective-risk social dilemmas, where the investment aims to avert the risk of losing more benefits due to climate change [20]. Among these existing frameworks, collective-risk social dilemmas have attracted most attention, both theoretically and experimentally [21–24]. In this simple, paradigmatic setup, subjects are divided into groups and repeatedly make decisions of investment with a target goal in mind, that represents the minimum amount the group needs to invest to avoid the undesired outcome. In the present context, achieving the goal means that the tragedy of commons such as dangerous climate change impact could be mitigated, otherwise the remaining individual wealth is at stake and can be completely lost with a certain loss probability.

We then argue what are effective strategies to achieve the target sum in collective-risk social dilemmas. Previous studies have shown that punishment and reward may help enhance cooperation and compliance[25–28]. A recent research [20] found particularly interesting results: the higher the risk of losing the accumulated earnings is, the easier it is to reach the collective target sum. Thus climate change mitigation is more likely to be achieved when the probability of mid- and long-term climate impact is higher. Besides, free-riders or non-cooperative players apparently have negative impact in achieving a collective goal. So punishment to free-riders may be an effective way to enhance the cooperation. Importantly enough, such a strategic change in collective dilemmas can be mapped to actual policies as recently discussed [13, 29, 30]. When the enhanced cooperation has outweighed the incurred cost of punishment, it is beneficial for players to achieve the goal. Such scenario has been observed in experiments with relatively long duration [31]. Nevertheless, by now the effect of punishment remains largely unclear [32–34]. For instance, punishment in short repeated experiments (typically 10 rounds or less) may enhance the cooperation but not the average payoff of the group [31, 35]. So in this particular situation costly punishment may be not helpful to improve the probability of achieving the target sum. Further, in such situation the total payoff could be even negatively correlated with the use of costly punishment [35]. We note, however, it is unknown whether above conclusions are also valid in a short repeated collective-risk dilemma game. To our knowledge it remains a challenge how to effectively achieve the target sum in a short repeated experiment with costly punishment.

In this paper we specifically investigate how costly punishment influences individual investment in the collective-risk climate dilemma game. To this end, we carry out a short (10 rounds) lab experiment where subjects were divided into independent groups of size M. Initially, all individuals had 20 monetary units (MU), and each subject was able to contribute 0, 1 or 2 MU to her group. If after 10 rounds the collective target of $10 \cdot M$ was achieved, players keep the money they saved. At variance with the traditional setup [20, 36], we introduced costly punishment: if there are free-riders — individuals that invest 0 MU — with a probability p, referred to henceforth as punishment risk, such subjects are fined with 3 MU. Moreover, punishment is not cost-free but has a cost of 1 MU that is evenly shared by all group members. Note that in this way, whether to punish free-riders or not is decided by punishing probability. Finally, if the group does not reach the target amount, with loss probability 0.5 (the loss probability is also 0.2 or 0.8 in some settings), all the individuals lose their savings, otherwise the remaining amount constitutes their earnings. Through such collective-risk game experiment, we are able to investigate behaviors of players in scenarios where deterrent of uncertain punishing risk is present.

We further notice that the effectiveness of punishment is accomplished through deterrence to players or subjective perception of players to the punishing risk. Nevertheless such subjective estimate could be different from the actual level of the risk [37]. For a very effective deterrence, there is a chess maxim telling that "the threat is stronger than its execution." Indeed, in some cases simulations have shown that making a threat of punishment can reduce the need to actually having to punish and improve cooperation [38]. To examine the efficiency of deterrence and how the punishment improve the cooperation through deterrence, we asked for the players, every time they acted as free-riders, whether they thought they will incur in a fine. Interestingly, we observe apparent irrational perception of the players who underestimate the risk to be punished. Besides, we also find that deterrence effectively reduces the number of free-riders as well as altruists. Our results show that costly punishment increases the likelihood to collect the desired amount, and thus play a key role in reaching the final goal. Besides, such finding is much more pronounced for very high punishment risk where deterrence is largely pronounced. We further uncover that the larger the group size is, the harder it is to accomplish the collective target for even large values of the punishment risk. These results point to the existence of a non-trivial tradeoff between enforcing measures and cooperative multi-country governance in climate change.

II. RESULTS

We first examine the efficiency of deterrence, i.e., whether the players respond rationally to a punishment risk in the

game. The results for the group size $M = 5$ are shown in Figure 1. In Figure 1(a) we present the variation of the estimated punishment ratio, q, measured as the ratio between the number of times subjects believed they will be punished and the number of times they opted to play as free-riders, as a function of the preset punishment risk p. As can be clearly seen, the perceived risk is always below the diagonal, i.e., that most of the players are not risk-averse, that is, most of the times that they did not contribute. They were willing to take the risk — conjecturing that they would not be fined. The biggest difference between the risk-perception q and the actual risk p happens at $p \approx 0.6$. By conducting this kind of irrational behavior, these players anticipate that their personal gains may be amplified. Such underestimate implies that the efficiency of deterrence is possibly weak in current experimental setup with random enforcement of punishment. We note, a previous study indicated that dynamically concentrated sanction of the punishment as an alternative may help improve the efficiency of deterrence [37]. In current study, however the punishment is possibly very effective only when it takes place with a high value of the punishment risk p. To verify this conjecture, in Figure 1(b) we quantify the extent of deterrence for different values of the punishment risk p, measured as the ratio between the number of times subjects opted to play as non-zero contributors to the target sum instead of free-riders in the presence of the punishment risk and the total number of tests at given punishment risk p. We observe that the extent of deterrence does not develop linearly with the punishment risk p. For small values of $p \leq 0.6$, the extent of deterrence is stable at around 0.7. Thus there are still an appreciable part of subjects who were willing to contribute none and took the risk. However, when $p \geq 0.8$, the ratio of this type of subjects decreases sharply, i.e., almost all subjects were effectively deterred to behave as none-zero contributors. Such scenario is simultaneously accompanied by the sharp decrease of the cost of the punishment in the game, as shown in Figure 1(c). We note, however, when the punishment risk $p \leq 0.6$, the cost of the punishment increases slightly for larger value of p. This discovery seems consistent with previous findings where the total payoff is negatively correlated with the use of costly punishment [35].

We then investigate how the cumulative investment in a game evolves with the number of rounds for different values of the punishment risk p. In Fig. 2 we show results in both games in which the final target was achieved (left panel) and games in which the final target was not achieved (right panel), as a function of p for groups of size $M = 5$. As can be seen in the figure, those games in which the final amount required was reached are dominated by a steady increase in the cumulative investments, without abrupt changes in the shape of the curves, for both values of p. In fact, the slope is roughly one, indicating that on average, individuals contributed 1MU per player in each round of the game. The behavior for the cases in which the final goal was not attained is however dependent on p. When the punishment risk p is very low, e.g., $p = 0.2$, we found that after 5 rounds the players tend to contribute even less in the presence of punishment risk. This implies that the deterrence is gradually relieved to the players. Nevertheless as the punishment risk increases, the average amount invested per round is higher steadily. Interestingly enough, even for high values of p, it is most of the times not enough. However, players do not stop contributing to the PGG, though they invest less and less as they approach the last round. Even if the probability of losing everything left at the end of 10 rounds is $1/2$, the latter behavior is rooted in the need to avoid additional loses that players might incur in if they act as free-riders: as fines are imposed with high probability, any eventual savings might be taken out by the fine itself or by the cost of applying it after 10 rounds.

We further investigate whether the above results are valid for different values of the group size M and how they affect the failure probability of a game's outcome. Figure 3 shows the cumulative investments averaged over all the members of the group and the failure probability, as a function of the size of the groups for four different values of the punishment risk p after 10 rounds. We also show in panel (c) the same results displayed in (b), but represented as a function of p for fixed values of the group sizes. The left (light blue) bars in each set of Figs. 3(a) and (b) display results for $p = 0$, that corresponds to the situation in which there is no punishment to free-riders. Two features are worth highlighting: the number of experiments for which the target amount was not achieved (failure probability) is remarkably high, which in turn increases with the size of the group, see also Figure 3c. This is a consequence of the low amount contributed to the PGG in all cases, 7.5785 ± 0.39002, 4.79167 ± 0.32773 and 4.025 ± 0.28662 for $M = 2$, $M = 5$ and $M = 10$, respectively. However, when the punishment (namely, $p > 0$) comes into play, the fraction of failures starts to drop, leading to an increase in the number of PGG in which the final target is reached. Interestingly, this decrease of the fraction of failures is significant only when the probability of being fined is large enough. Indeed, for $p \leq \frac{M(1-p*)}{3M+1}$ (where $p*$ is the loss probability, see Methods), the more rational strategy to maximize benefits is to free ride. As p increases beyond this bound, adopting a fair-share strategy is the best, as even a single defector would earn less. However, as it can be seen in the figure, only for high values of p (beyond $p = 0.8$ in our case), punishment has an impact on the players' behavior. This implies irrational perception of the players who initially underestimate the risk to be punished and is consistent with our previous findings presented in Figure 1. As for the dependence with the size of the group, the same pattern with respect to the case $p = 0$ is observed, as shown more explicitly in Figure 3c. The increase of average investment and the drop of failure probability both imply the deterrence due to the punishment is effective for different sizes of groups. Nevertheless, only when $p = 1$, the probability of failure is measured to be 0 for the largest group in our experiments ($M = 10$) and the average amount contributed increases with M (10 ± 0.12403, 10.15 ± 0.1163 and 10.25 ± 0.17813 for $M = 2$, $M = 5$ and $M = 10$, respectively). These findings hence suggest that the larger the size of the groups is, the higher the punishment to free riders should be for the final goal to be attained. Further, the results above seem robust against different values of the loss probability (Figure 4).

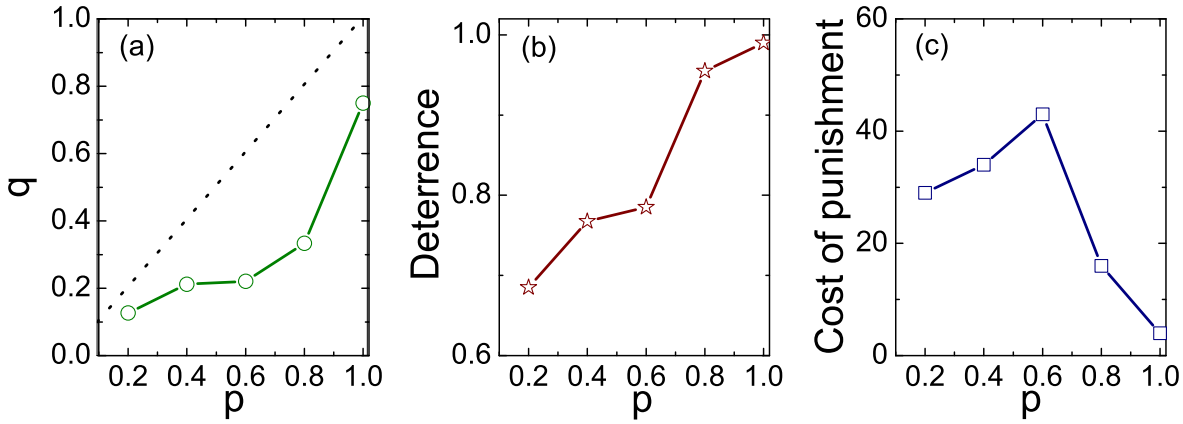

Fig. 1. Dependence of (a) the risk-perception q, (b) deterrence, and (c) cost of the punishment on the punishment risk p. In (a) we show that the perception that the risk to be punished, q, is lower than the likelihood of punishment p. The quantity of q is measured as the ratio between the number of times a player thought she was going to be punished and the number of times the same subject played as free rider. In (b) the extent of deterrence is measured as the ratio between the number of times subjects opted to play as non-zero contributors to the target sum in the presence of the punishment risk and the total number of tests at given punishment risk p.

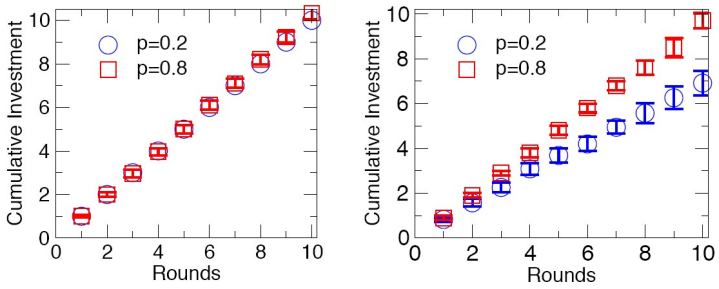

Fig. 2. Evolution of the cumulative investment with the number of rounds. The two panels show the cumulative amounts contributed to the PGG as a function of the number of rounds played for groups of size $M = 5$ and two values of the punishment risk p. The left panel displays how this quantity varies when considering only the games in which the final target was achieved, whereas the right panel shows results averaged over games in which it was not. Interestingly, even in the case in which the target was almost unreachable (left panel, $p = 0.2$), the players kept donating, which is a consequence of the punishment mechanism and that the loss probability was set to 0.5. Error bars represent the Standard Error of the Mean (SEM).

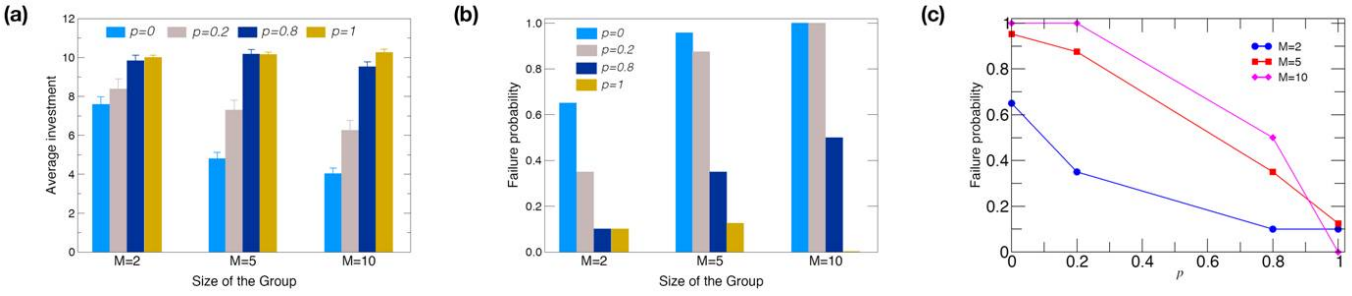

Fig. 3. Average investment and failure probability as a function of the group size M and p. Panels (a) and (b) shows both quantities for four different values of the punishment risk p, while panel (c) represents the fraction of failures as a function of p and M. In (a) and (b), and for all groups of histograms, the values of the punishment risk p are as indicated. We found that punishment is effective only for high values of p, meaning that free-riders will be fined with a high probability. The dependence with the group size shows that for all values of $p < 1$, the larger the size of the group is, the less amount is contributed and the harder it is to achieve the final goal. Error bars represent the Standard Error of the Mean (SEM).

It is interesting to observe the behavior of failure rate for different values of the loss probability $p*$ while keeping small punishment risk $p = 0.2$ among panels (a-c) in Figure 4. For example, as $p*$ grows from 0.2 to 0.5, it is shown that the failure rate decreases from 0.6 to 0.35 with a minor increase of average investment. We believe that the key contributors of such behavior are free-riders in the group. Our argument is the following. The free-riders usually gives a higher priority on improving their own earnings than on common interests. Nevertheless when both p and $p*$ are small, even free-riders would feel the danger of loss is small, and the possible loss may be largely avoided by cherishing common interests when the result is on the margin. As long as such strategy works, it would effectively decrease the failure rate of groups and has much less effect on the average investment when $p*$ increases. As shown in panels (a-b), the strategy still holds when $p*$ mounts to 0.5. However when $p*$ continues growing, free-riders gradually feel the danger of loss is high enough thus for better personal benefits they focus more on their own earnings instead of common interests. As a result, we observe a high failure rate of 0.3 with $p* = 0.8$ in panel (c) although high $p*$ enhances noticeably higher investment from others.

Next, we analyze individual behavior. We consider three different possibilities: i) selfishness, typifying free riders that contribute 0 to the PGG but obtain the largest benefit if the global target is achieved; ii) fairness, characterizing those individuals contributing the fair-share of 1MU; and iii) altruism, describing the behavior of those individuals that contribute the most (2MU). Figure 5 shows the distribution of the three behaviors as a function of the punishment risk p for different group sizes. As it can be seen, regardless of the group size, the number of free-riders decreases in general when the punishment risk is nonzero and grows. And this trend is more apparent when players mostly believe they will be punished with a high punishment risk p. These results indicate that the deterrence comes into play for $p > 0$ and forces free riders to behave as fair-share investors to avoid punishments. The opposite trend is observed concerning the number of subjects contributing the fair-share. Interestingly enough, the deterrent effect of punishment makes altruistic contributions to decrease as well. Admittedly, the selfish and altruistic investments go hand-to-hand, namely, as soon as the number of free-riders decreases due to the higher values of the punishment risk, the number of maximal contributions does not remain constant but also decreases in favor of the fair-share behavior. We argue that the altruists may have realized a bigger probability of the goal achieved, thus strategically choose to contribute less while not affecting the outcome. Indeed, as seen in Figure 2 for the games in which the final goal was achieved, there is no abrupt change in the amounts contributed in each round of those games, which is a further indication that the fair-share strategy is quite stable as rounds go by. And for different group sizes, we do not observe significant variations of the previous patterns, except for the largest size $M = 10$ and $p = 1$, a scenario in which almost all players (87%) contribute the fair amount. Note that the latter case shows the lowest number of free-riders but also of altruism level.

III. DISCUSSION

The results of the present collective-risk social dilemmas experiments have important implications. Even if our experimental setup does not capture all the complexity of a collective governance problem such as agreeing on measures to mitigate dangerous climate changes, it certainly gives further insights into a class of dilemmas -the tragedy of commons [39]- that can provide hints to interpret and shape the dynamics in dilemmas. Our findings show that punishment accomplished through appreciable deterrence could be an effective mechanism to achieve global targets in the current context. At the same time, however, we have shown that in order for such an enforcing measure to be efficacious, it should be perceived as almost certain, otherwise its effects might be blurred. Hence, if we realize that individuals, institutions or the private sector are hesitant to provide a collective good without being enforced because the short-term benefits of defection are higher, then it follows that international treaties should necessarily compel governments to adopt measures aimed at overcoming those short-term incentives to free ride.

In conclusion, and with all due caution, the present study suggests that in climate negotiations, measures such as imposing economic sanctions to non-cooperative countries might be effective. The results from our collective-risk games are robust against different group sizes and/or different loss probabilities. With appreciable deterrence the behaviors of all players become convergent as fair-share investors while maximizing the probability to achieve the target sum.

Compared to the previous result available [20], our results are not very different. Specifically, Ref.[20] reported, for groups of size six, that with a loss probability of $1/2$ only 1 out of 10 groups reached the target. That is the same result obtained in our experiment for $M = 5$ and $p = 0$, the closest comparable setup.

IV. METHODS

a. **Maximizing payoffs.** As mentioned before, even with punishment in place, free-riding could be the rational strategy to maximize benefits depending on the value of the punishment risk p and of the loss probability p^*. To see this, let us assume that all players behave in the same way. In one scenario, where all players are free-riders, the target will never be reached. In the second scenario, where all players adopt a fair-share strategy, the target will always be attained. Thus, when is it more profitable in terms of the likelihood to maximize benefits to play as a free-rider or as a fair-sharer? The final expected payoff of the free-riders in the first situation would be:

$$\Pi^f_{(1)} = \Pi^i - \sum_{j=1}^{N}(3p + \frac{p}{M}),$$

while in scenario two it would be:

$$\Pi^f_{(2)} = \Pi^i - \sum_{j=1}^{N} x,$$

where $\Pi^f_{(\cdot)}$ and Π^i are, respectively, the final expected payoff and the initial capital, M is the size of the group, N is the number of rounds played, and x is the contribution to the common pool. Thus, for defection to be better than fair-share, the relation $\Pi^f_{(1)} \geq \Pi^f_{(2)}$ should be verified, which leads to the condition (setting $x = 1$):

$$p \leq \frac{M}{3M+1},$$

As free-riders only collect their benefits with probability $(1 - p*)$, the final condition for p is

$$p \leq \frac{M(1-p*)}{3M+1}.$$

b. **Experimental sessions.** The experiments of collective-risk social dilemmas were conducted from July of 2013 to November of 2013 and they involved a total of 720 freshmen and sophomore (coming from different majors) at Wenzhou University, China. Subjects consent was obtained before starting the experiments and after they answered a questionnaire. Each experimental session required the simultaneous participation of 20 subjects, who were randomly divided into several groups (namely, the setup was completely anonymous). Both the size M and composition of the groups were kept constant during the whole experiments. Within each group, subjects repeatedly played 10 independent rounds of the game. Each subject started with an initial endowment of 20 monetary units (MU). Each experimental round consisted of the following steps:

- At each round, all the subjects were asked simultaneously whether they would independently contribute 0 MU, 1 MU, or 2 MU to the climate account.
- After taking their investment decisions, every subject was shown the following information during 30 seconds: (i) individual contribution (0 MU, 1 MU, or 2 MU); (ii) the collective investment of his/her group in the current round; (iii) the remaining gap between the cumulative contribution and the required target sum of the group.

Similar to previous experimental setups [20], the total investment required for one group of size M to reach its target was set to $10 \cdot M$ (equivalent to 1 MU per subject per round on average). If the overall contribution after 10 rounds was equal or greater than the collective target, individuals could keep the money saved. On the contrary, if the target sum was not reached, subjects could lose all their savings with a loss probability that we set to 0.5 in most of the sessions carried out.

Based on the above-mentioned basic setup of the collective-risk dilemmas game, we introduced costly punishment into the experiments. Distinguishing from previous theoretical researches about pool- or peer-punishment in game theory [33], the implementation of punishment in the present work is directly related to the performance of the group, and the cost of imposing a fine to non-cooperators was evenly distributed among all group members, regardless of their investment behavior. More specifically, if at any round of the game there exist non-cooperators -people that contribute zero to the PGG. To avoid the problem of second order free riders, the decision of the punishment is made based on the environment which is controlled by the probability p. The cost of punishment is paid in the form of tax, implying participants may have to pay the cost even in the case of punishing themselves. Thus, a fine of 3 MU is imposed to all those selfish players in the group who behave as free-riders with the probability p. At the same time, the total cost of punishment, namely, 1 MU per selfish player -since the cost of 1 MU is associated to each fine applied- is equally distributed among the M players of the group. In addition, there is no punishment with the probability of $1 - p$. Such kind of punishment can be called an "imperfect" punishment.

Finally, at the end of the experimental session, the remaining monetary units (MU) were changed into real money. Earnings -including the show-up fee- ranged from 20¥ to 40¥ and the conversion rate applied was 1 MU = 1¥. The value of the show-up fee in our experiment corresponds to the minimum earnings. Here, the earnings of every participant include a fixed base salary and a floating commission. Every participant could receive the base salary as long as he/she has shown up. Whereas the commission received depends on his/her strategy as well as his/her partners'. The earnings of participants are thus fluctuating while the minimum is the base salary. Altogether, the results reported here come from 232 groups (120 groups of size $M = 2$, 64 of size $M = 5$ and 16 of size $M = 10$). The instructions and questionnaires took 5 to 10 minutes and the entire game took 30 to 35 minutes for the 10 rounds. The average earning of all the participants was 32.2 ¥.

ACKNOWLEDGMENTS

We would like to thank K.-Z. Jin, C. Gracia-Lázaro and A. Sánchez for helpful discussions. This work was supported by the National Natural Science Foundation of China (Grants 61203145), by the Natural Science Foundation of Zhejiang Province (Grant No. LY17F030005), and by the Youth Foundation of Social Science and Humanity, Ministry of Education of China (Grant No. 20YJCZH077). M.P. was supported by the Slovenian Research Agency (Grant Nos. P1-0403 and J1-2457).

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

[1] B. C. O'Neill and M. Oppenheimer, *Science* **296**, 1971–1972 (2002).
[2] S. Manabe and R. J. Stouffer, *Science* **364**, 215–218 (1993).
[3] W. S. Broecker, *Science* **278**, 1582–1588 (1998).
[4] J. E. Hansen, *Climate Change* **68**, 269–279 (2005).
[5] Intergovernmental Panel on Climate Change, *Climate Change* (2014) (Synthesis Report, Accessible at *http://www.ipcc.ch/report/ar5/syr/*).
[6] G. P. Peters, R. M. Andrew, T. Boden, J. G. Canadell, and P. Ciais, *et al. Nature Clim. Change* **3**, 4–6 (2013).
[7] D. J. Griggs and M. Noguer, *Weather* **57**, 267–269 (2002).
[8] S. Barret, *Environment and Statecraft.* (Oxford University Press, 2003).
[9] J. Hansen, M. Sato, P. Kharecha, D. Beerling, R. Berner, *et al*. *Open Atmospheric Science Journal* **2**, 217–231 (2008).
[10] R. B. Alley, J. Marotzke, W. D. Nordhaus, J. T. Overpeck, and D. M. Peteet, *et al*., *Science* **299**, 2005–2010 (2003).
[11] M. Finus, *Int. Rev. Environ. Res. Econ.* **2**, 29–67 (2008).
[12] J. Heitzig, K. Lessmann, and Y. Zou, *Proc. Natl Acad. Sci. USA* **38**, 15739–15744 (2011).
[13] P. M. Regan, *The politics of global climate change*. (Taylor and Francis, 2015).
[14] M. M. Bechtel and K. F. Scheve, *Proc. Natl Acad. Sci. USA* **110**, 13763–13768 (2013).
[15] T. Dietz and J. Zhao, *Proc. Natl Acad. Sci. USA* **108**, 15671–15672 (2011).
[16] M. Milinski, D. Semmann, H. J. Krambeck,and J. Marotzke, *Proc. Natl Acad. Sci. USA* **103**, 3994–3998 (2006).
[17] Y. M. Svirezhev, B. Werner, and H. J. Schellnhuber, *Environmental Modeling & Assessment* **4**, 235–242 (1999).
[18] S. Barrett and A. Dannenberg, *Proc. Natl Acad. Sci. USA* **109**, 17372–17376 (2012).
[19] P. K. Dutta and R. Radner, *Proc. Natl Acad. Sci. USA* **101**, 5174–5179 (2004).
[20] M. Milinski, R. D. Sommerfeld, H. J. Krambeck, F. A. Reed, and J. Marotzke, *Proc. Natl Acad. Sci. USA* **105**, 2291–2294 (2008).
[21] V. V. Vasconcelos, F. C. Santos, and J. M. Pacheco, *Nature Clim. Change* **3**, 797–801 (2013).
[22] X. J. Chen, A. Szolnoki, and M. Perc, *Phys. Rev. E* **86**, 036101 (2012).
[23] C. Hilbe, M. A. Chakra, P. M. Altrock, and A. Traulsen, *PLoS ONE* **8**, e66490 (2013).
[24] A. R. Góis, F. P. Santos, J. M. Pacheco, and F. C. Santos, *Sci. Rep.* **9**, 16193 (2019).
[25] M. H. Duong and T. A. Han, *Proc. R. Soc. A* **477** 20210568 (2021).
[26] T. A. Han, *J. R. Soc. Interfac* **19** 20220036 (2022).
[27] K. Li Y. Mao, Z. Wei, R Cong, *Chaos, Solitons and Fractals* **143** (2021) 110591.
[28] L. Chen, J. Sun, K. Li, and Q. Liang, *Physica A* **591** (2022) 126804.
[29] S. Barrett, *Oxford Rev. Eco. Policy* **24** 239–258 (2008).
[30] S. Barrett and M. Toman, *Global Policy* **1**, 64–74 (2010).
[31] S. Gächter, E. Renner, and M. Sefton, *Science* **322**, 1510–1510 (2008).
[32] C. Hauert, A. Traulsen, H. Brandt, M. A. Nowak, and K. Sigmund, *Science* **316**, 1905–1907 (2007).
[33] A. Traulsen, T. Röhl, and M. Milinski, *Proc. R. Soc. B* **279**, 3716–721 (2012).
[34] S. Barrett, *Towards a better climate treaty* (Fondazione Eni Enrico Mattei Venice, 2002).
[35] A. Dreber, D. G. Rand, D. Fudenberg, and M. A. Nowak, *Nature* **452**, 348–351 (2008).
[36] J. Jacquet, K. Hagel, C. Hauert, J. Marotzke, and T. Röhl, *et al*., *Nature Clim. Change* **3**, 1025–1028 (2013).
[37] M. Kleiman and B. Kilmer, *Proc. Natl Acad. Sci. USA* **106**, 14230–14235 (2009).
[38] T. Cimpeanu and T. A. Han, *2020 IEEE Congress on Evolutionary Computation (CEC). IEEE* 1–8 (2020).
[39] G. Hardin, *Science* **162**, 1243–1248 (1968).

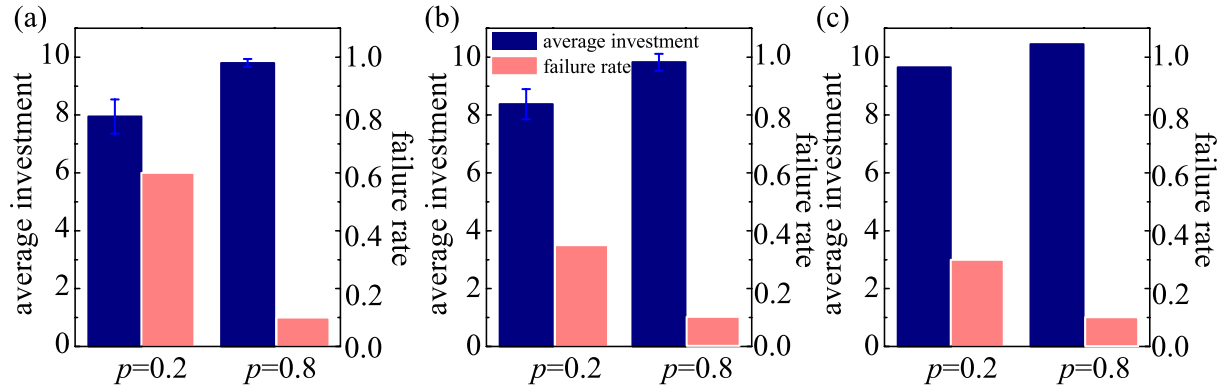

Fig. 4. Average investment of subjects and failure rate of groups during 10 rounds of collective-risk dilemma game depending on punishment risk p for different loss probability $p*$. The values of loss probability risk $p*$ are 0.2 (panel a), 0.5 (panel b) and 0.8 (panel c).

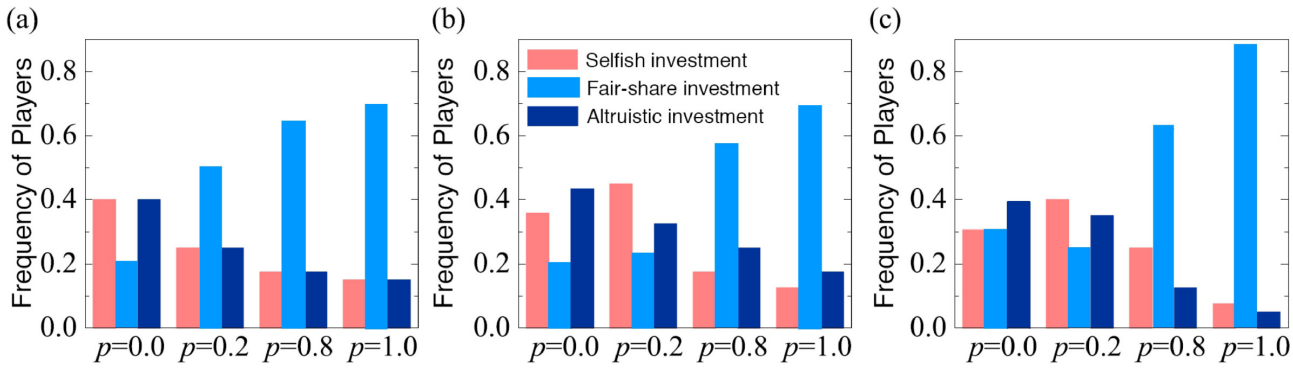

Fig. 5. The distribution of classes of players as a function of the punishment risk p for different group sizes M. Increasing the punishment risk reduces the number of free-riders in the game, but at the expense of a decrease in the number of maximal contributors. As a result, most of the players adopt a fair-share strategy, which however is not enough to reach the final target in many of the games, as indicated by the fraction of failures in Fig 3. The sizes of the groups are, from (a) to (c), $M = 2$, $M = 5$ and $M = 10$, respectively.