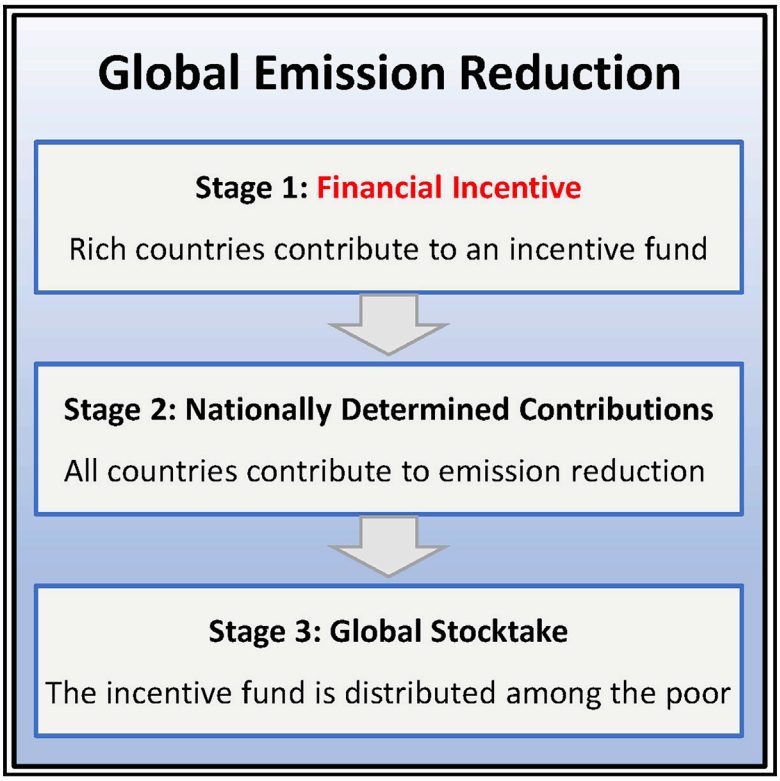
One Earth

Article

Financial incentives to poor countries promote net emissions reductions in multilateral climate agreements

Graphical abstract



Highlights

- Global emissions reduction can be achieved in the absence of any binding enforcement
- The rich incentivize the poor through FIs, and the poor increase their ER accordingly
- Overall,FIslead not only to higher ER, but also to higher socialwelfare
- The results are robust to changes of different sets of parameters

Authors

Yali Dong, Shuangmei Ma, Boyu Zhang, Wen-Xu Wang, Jorge M. Pacheco

Correspondence

zhangby@bnu.edu.cn (B.Z.), wenxuwang@bnu.edu.cn (W.-X.W.), jmpacheco@math.uminho.pt (J.M.P.)

In brief

The Paris Agreement established the latest ruling in dealing with climate change. However, the effectiveness of nationally determined contributions and financial mechanisms in the agreement is unclear. Dong et al. study the global emissions reduction problem through behavioral experiments and game theory. They find that financial incentives systematically increase emissions reduction and social welfare in the absence of any binding enforcement mechanisms, where most mitigation actually stems from the developing countries, while the developed countries mostly incentivize their actions.

Dong et al., 2021, One Earth *4*, 1141–1149
August 20, 2021
https://doi.org/10.1016/j.oneear.2021.07.006

CellPress

One Earth

Gold Standard for the Global Goals | 3 tons of CO_2 offset

CellPress OPEN ACCESS

Article

Financial incentives to poor countries promote net emissions reductions in multilateral climate agreements

Yali Dong,[1,6] Shuangmei Ma,[1] Boyu Zhang,[2,6,7,*] Wen-Xu Wang,[1,6,*] and Jorge M. Pacheco[3,4,5,6,*]

[1]School of Systems Science, Beijing Normal University, Beijing 100875, PRC

[2]Laboratory of Mathematics and Complex Systems, Ministry of Education, School of Mathematical Sciences, Beijing Normal University, Beijing 100875, PRC

[3]Centro de Biologia Molecular e Ambiental, Universidade do Minho, 4710-057 Braga, Portugal

[4]Departamento de Matemática e Aplicações, Universidade do Minho, 4710-057 Braga, Portugal

[5]ATP-Group, P-2744-016 Porto Salvo, Portugal

[6]These authors contributed equally

[7]Lead contact

*Correspondence: zhangby@bnu.edu.cn (B.Z.), wenxuwang@bnu.edu.cn (W.-X.W.), jmpacheco@math.uminho.pt (J.M.P.)

https://doi.org/10.1016/j.oneear.2021.07.006

SCIENCE FOR SOCIETY Cooperation to reduce greenhouse-gas emissions has profound impacts on sustainable development. We explore the impact on global cooperation of a financial incentive (FI) by which developed countries can incentivize developing countries to reduce carbon emissions. We perform both behavioral experiments and game theoretical analyses to clarify possible result bias stemming from bounded rationality of humans or from inherent selfish motivations among different countries. Our results show that FIs systematically contribute both to increase global contribution in emissions reduction and to effectively reduce emissions globally, significantly improving the probability of achieving the targets specified in international agreements. This happens in the absence of any binding enforcement, and strongly suggests that developed countries should divert some of their financial resources to incentivize developing countries to reduce their emissions.

SUMMARY

Reducing global greenhouse-gas emissions needs global cooperation and will have a positive and profound impact on sustainable development. Climate agreements, in line with the UNFCCC, encourage developed countries to provide funds to help developing countries adapt and mitigate. However, up to now, no financial incentive (FI) has been implemented, and it remains unclear to what extent FIs can increase net emissions reductions (ER). Here we investigate a restrictive form of FI, employing both behavioral experiments and game theoretical analysis. We show that FIs significantly increase both ER and social welfare in the absence of any binding enforcement. We also find that the more developed countries invest in FIs, the more developing countries mitigate. This induces developed countries to incentivize developing countries to adapt and mitigate via FIs, resulting in a net global increase in ER. Our results are robust to different monitoring periods, loss probabilities, and mitigation cost ratios.

INTRODUCTION

The Paris Agreement (PA), signed in 2015, resulted in an inclusive, binding treaty that succeeds the Kyoto Protocol and the Copenhagen Accord. It has been argued that the PA constitutes a significant breakthrough in international climate negotiations,[1–4] reinstating the United Nations Framework Convention on Climate Change (UNFCCC) as a forum for dynamic multilateralism.[5,6]

At present, the Intergovernmental Panel on Climate Change (IPCC) has pointed out that the way to solve the threats posed by climate change is to reduce emissions by at least 50% of the 2000 level by 2050.[1–4,7] To achieve this goal, widespread cooperation is required. Given that countries are heterogeneous in terms of wealth (according to the World Bank, countries are categorized as high income, middle income, and low income),[8] emissions reduction (ER) costs (the World Bank uses the GDP per kilogram of CO_2 emissions as a surrogate for ER cost and,

thus, high-income countries have higher costs than do middle- and low-income countries)[9] and risks (according to the IPCC, middle- and low-income countries are more vulnerable to extreme events and disasters than high-income countries)[10] differ when countries face climate disaster. Given the disparities between high-income and middle- and low-income countries, cooperation on ER constitutes a highly non-trivial problem, and more so when we take into consideration that global cooperation must be achieved through international agreements whereby sanctioning mechanisms are very difficult to implement. Consequently, here we explore the impact of implementing a financial incentive (FI) mechanism inspired by what the UNFCCC designates as ''Financial Mechanism'' (FM), whose operating entities were included as an important aspect of the PA (see Note S1.2 for detailed discussions).[1,11,12] According to the FM in the PA, developed country parties shall provide financial resources to assist developing country parties, and in our experimental design, subjects with high endowment can incentivize subjects with low endowment to contribute to ER through an FI, which mimics the FM in the PA. Unlike alternative mechanisms already proposed,[13–22] such as cap and trade and a carbon tax with punishment for over-emission, the FIs discussed below do not require any binding enforcement mechanism of control or any penalties to discourage free riding. Thus, a natural question is whether FIs can help to promote ER.

Following previous studies, the problem of cooperation to ensure ER may be framed as a multi-period threshold public goods dilemma game involving players with different amounts of wealth, mitigation costs, and (non-negligible) risks of future losses.[23–40] In this game, players are challenged to reach a pre-defined group target evaluated at the end of a series of periodical contributions; if the target is not met, players will lose their wealth with pre-defined probabilities. In this work, we set up groups of six individuals,[24,26,28,38] in which we simulate the global ER problem and investigate the impact on ER in the presence and in the absence of FIs.

In each group, we randomly select a player to represent a high-income country (rich), while the remaining five players represent middle- and low-income countries (poor). Rich players are not only wealthier (they start with higher initial endowments) but also may have higher ER costs compared with poor players: indeed, in 2014, the GDP per kilogram of CO_2 emissions—used as a surrogate for cost of ER—was 3.2 times higher for high-income countries compared with middle- and low-income countries.[9] It is important to point out that, in each group, only one participant assumes the role of a high-income country. This way, the interaction between ''rich'' countries is not explicitly included in each group, unlike what happens with players acting as low-income countries (see Note S6, where interactions among the rich in larger groups are incorporated).

Two different game settings were implemented, here designated as Control and Treatment. In Control, there is no FI. Both rich and poor subjects can contribute to ER in each period (ER stage). At the end of the T periods, the total contributions are compared with the target announced (quantitative details in the experimental procedures). In Treatment, each of the T periods comprises an additional FI stage before the ER stage, in which the rich subject in the group can contribute to a fund whose resources are distributed by the poor subjects of the group after their contribution is made in the ER stage. Naturally, several possibilities exist to allocate the resources in the fund. Here, we decided to allocate the resources proportional to the contribution of the poor to ER (no contribution means no transfer, see Note S1 for more details). Thus, under the FI, the rich are allowed to contribute to the fund but not to benefit from it, while the poor benefit from the fund without contributing to it. In the following we shall designate the aforementioned fund also by FI, as it is associated with the FI stage.

In our experimental design (as well as in the theoretical analysis), endowments are finite and limited and the transfers are both endogenous and budget balanced. To this end we further impose that resources allocated by the rich to ER cannot be used for FIs and vice versa. Importantly, poor subjects cannot contribute more to ER even when their wealth increases via FIs (see the experimental procedures for details). All these restrictions, as discussed below, impose stringent conditions on the FIs implemented here.

In addition to inequality, our experimental design attempts to mimic the global stocktake of the PA, whereby each country's ER will be monitored every 5 years, starting in 2023,[1–4] such that future decisions may be contingent on perceived status. As global emissions should decrease to 50% of the present level by 2050,[1–4,7] this means approximately six monitoring periods.

We performed behavioral experiments with the multi-period threshold public goods dilemma game just described involving 840 volunteer undergraduate and graduate students recruited from Beijing Normal University who had not taken classes on game theory and economics. A rich subject has an initial endowment five times larger than a poor subject. We considered two kinds of ER cost factors: s for a rich subject, where s = 2 and s = 3 (we always have s = 1 for every poor subject), and two values for the total number of monitoring periods, T = 6 and T = 10. At the end of each multi-period experiment, if the ER target was not reached, then rich and poor subjects lost all their savings with probabilities r_R and r_P, respectively (so-called risks of failure;[23,24,26–29,33,34,36–38,40,41] we allow r_R and r_P to take the values 0.5 and 0.7, see below).

Twelve experimental sessions were performed: six Control sessions and six Treatment sessions. Each session, associated with specific values of T, s, r_R, and r_P, was repeated employing 10 different groups of six individuals. Specifically, $(T, s, r_R, r_P) = (6, 2, 0.5, 0.5)$, $(6, 3, 0.5, 0.5)$, $(6, 3, 0.5, 0.7)$, $(6, 3, 0.7, 0.7)$, $(10, 2, 0.5, 0.5)$, and $(10, 3, 0.5, 0.5)$ in Controls 1–6, respectively, and the parameter settings in Treatments 1–6 were the same as those of Controls 1–6. In line with previous experiments,[23,24,28,29,33,37,38,42] subjects had limited options in both FI and ER stages; in the experimental procedures we explain in detail all options available, while in Note S1 we provide details of the experimental procedures.

RESULTS

Success rates and relative ER

We start by investigating possible scenarios that may result from the experiments by carrying out a game theoretical analysis of (some of) the subgame perfect Nash equilibria (SPNE; which is a refinement of Nash equilibria for a multi-period game) stemming from modeling the experimental setting, as well as an

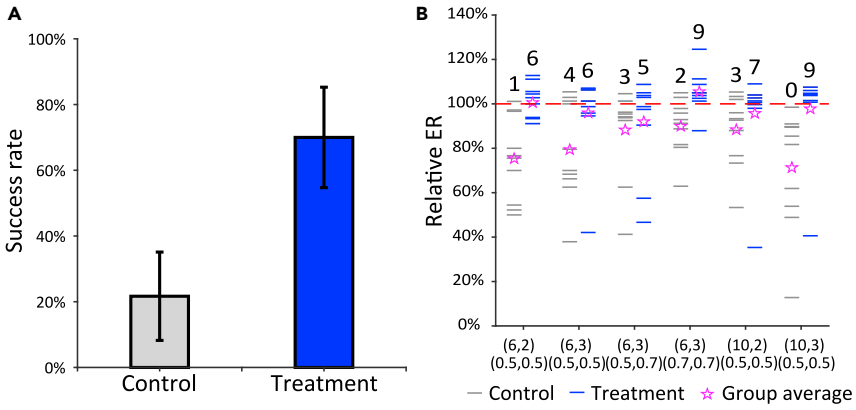


Figure 1. Success rates in meeting the ER targets and relative ER at the group level

(A) Aggregated data stemming from all the experiments. Data are presented as the mean ± standard deviation (defined as $\sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})^2}$), where $\{x_1, x_2, \ldots, x_N\}$ are the observed values, $\bar{x}$is the mean value of these observations, and N is the sample size). The success rates for Control and Treatment are 0.217 (gray bar) and 0.7 (blue bar), respectively, and the standard deviations are 0.134 (black error bar) and 0.153 (black error bar), respectively. Overall, FI leads to a marked increase in the success rate.

(B) The ER group achievement relative to the target ER value (in each column, there are 10 lines, one per group, gray lines for Control and blue lines for Treatment, whereas the pink star indicates the average over the 10 groups). To reach the target, the relative ER needs to reach 100% (red dashed line). The number above each column indicates the number of groups (of a total of 10) that reached the target.

identification of their stability through an evolutionary game theoretical analysis (where subjects are assumed bounded rational and learn to play the game)[43] (full details in the experimental procedures and Notes S2–S4). Given the multi-period structure of this game, individual strategies can be very complicated, and the existence of asymmetric SPNE is possible. In the following, we consider a specific subset of SPNE that we designate as quasi-symmetric SPNE (QSPNE), at which equilibrium all poor subjects use the same strategy. Furthermore, we focus on the sustainment of cooperation through the GRIM strategy, where a subject using GRIM will contribute to ER or FI in period t only if the total ER or FI in the previous periods in not less than a pre-defined value. We emphasize that the purpose of the theoretical analysis is not to perform a full analysis of the game, but to provide intuition on whether and how FI works.

We find that, in the Control case, a selfish, non-cooperative SPNE dominates for $r_R = r_P = 0.5$, at which SPNE no one contributes to ER (Figure S1). In the Treatment case, if the contribution of the rich to FIs positively correlates with the contribution of the poor to ER, then there is a narrow basin of attraction toward a set of (stable) cooperative QSPNE (which we designate as incentive QSPNE) in which only the poor contribute to ER, while the rich contribute solely to FIs (Figure S2). At those incentive QSPNE, both rich and poor subjects have higher wealth compared with the selfish SPNE even if, similar to the experimental design, poor subjects cannot contribute more to ER when their wealth increases via FIs (see the experimental procedures). Moreover, higher contributions by the rich to FIs are predicted to lead to higher investments in ER by the poor, but up to a limit, beyond which the incentive QSPNE become unstable.

We now focus on the behavioral experiments, whose results will also enable us to confirm or dismiss the theoretical predictions made above. Figures 1A and 1B show the success rates (fraction of groups that reach the target) and relative ER (group average ER relative to the target ER value), respectively, obtained in the experiments (for further details of relative ER at group level see Figure S3). Compared with Control (data in gray, ER stage only, per period), FIs lead to systematic improvements in success rates in Treatment (data in blue, FI stage + ER stage, per period). Furthermore, the total ER in Control is significantly below the target, whereas in Treatment it is not statistically different from the target (see also Table S1). These results indicate that FIs significantly promote ER in experimental scenarios involving diverse numbers of periods, cost factors, and risk combinations.

Behavior of rich and poor subjects

To uncover the effect of FIs in ER, we now explore the relative contributions by both rich and poor subjects, defined as their total contributions divided by their initial endowments. Figures 2A and 2B show relative contributions in Control and Treatment, respectively, where the contributions are by poor subjects to ER and by rich subjects to both ER and FIs (for further details of relative ER at group level see Figure S3).

We find that poor subjects contribute significantly more to ER in Treatment than in Control, despite being unable to use the tokens received from the rich via FIs to ER. Importantly, our experiments confirm that the investment made by poor subjects is positively correlated with the amount contributed to FIs by the rich at the group level (Pearson correlation coefficient = 0.7986, $p < 0.001$), which, in view of our theoretical analysis, suggests that the incentive Nash equilibria (NE) may be stable (see Figure S4 for further analysis of the correlation between ER by the poor and FI by the rich at the group level). This indicates that the availability of FI plays a key role in enhancing both the total ER achieved and the success rates in reaching the ER target. From the outset, it was not clear whether the rich would contribute to FIs, whether the poor would actually invest additionally in ER in the presence of FIs, or whether the additional investment, if made, would be sufficient. Our results indicate a statistically significant yes to all three questions. As a result, the rich mostly contribute to FIs, naturally implying (given the nature of the game at stake) that ER from the rich decreases in Treatment, compared with Control.

Figure 3 aggregates the information contained in Figure 2, at the same time separating successful from failed groups. Naturally, in Control, both rich and poor subjects invest more in ER

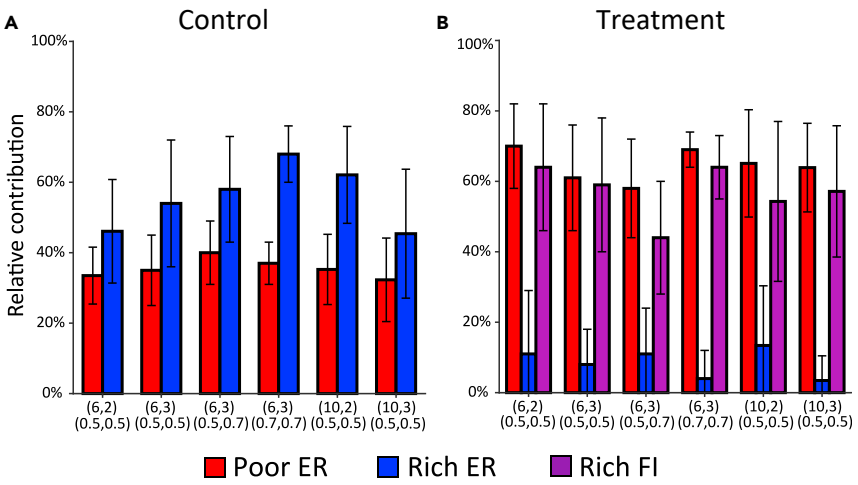


Figure 2. Average relative contributions (e.g., contributions relative to initial endowments) at the group level

Data are presented as mean ± standard deviation.

(A) In Control, relative contributions to ER of rich subjects (blue bars) are significantly higher than those of poor subjects (red bars) (Mann-Whitney U test, $p < 0.01$). The average relative contributions for poor and rich subjects in Control experiments 1–6 are 0.335, 0.35, 0.4, 0.37, 0.353, and 0.323 and 0.461, 0.54, 0.58, 0.68, 0.621, and 0.454, respectively, and the standard deviations (black error bars) are 0.081, 0.1, 0.09, 0.06, 0.1, and 0.119 and 0.147, 0.18, 0.15, 0.08, 0.138, and 0.183, respectively (see Table S1).

(B) In Treatment, poor subjects contribute significantly more to ER than rich subjects (Mann-Whitney U test, $p < 0.01$), who now opt to transfer wealth to poor subjects via FIs. The average relative contributions for poor subjects to ER (red bars) and rich subjects to ER (blue bars) and FIs (purple bars) in Treatment experiments 1–6 are 0.7, 0.61, 0.58, 0.69, 0.651, and 0.639; 0.11, 0.08, 0.11, 0.04, 0.134, and 0.035; and 0.64, 0.59, 0.44, 0.64, 0.543, and 0.572, respectively (see Table S1). Importantly, the wealth from FIs cannot be used by the poor to mitigate (see experimental procedures). Furthermore, the standard deviations (black error bars) of the relative contributions for poor subjects to ER and rich subjects to ER and FIs in Treatment experiments 1–6 are 0.12, 0.15, 0.14, 0.05, 0.152, and 0.126; 0.18, 0.1, 0.13, 0.08, 0.169, and 0.07; and 0.18, 0.19, 0.16, 0.09, 0.227, and 0.186, respectively. The data were analyzed at the group level to avoid interdependence of outcomes for members of a given group.

in the successful groups than in the failed groups. In Treatment, we see that in successful groups, contributions to ER by poor subjects and to FIs by rich subjects are both significantly higher compared with failed groups. However, rich subjects in successful groups contribute significantly less to ER than in failed groups. This, again, suggests that FIs play a significant role in stimulating poor subjects to mitigate. In failed groups, insufficient contributions by the rich to FIs induce small contributions to ER by poor subjects, leading to failures. Taken together, our results suggest that, under the conditions defined in the experiment, rich countries ought to divert most of their contributions from ER to FIs, since this not only enhances the chance of collective success, but also leaves all group members wealthier compared with Control. These features rely on the positive correlation between contributions to FIs by the rich and investment in ER by the poor.

Time evolutions of individual behaviors

In Figure 4 we show the results of a longitudinal analysis of individual behavior portraying the dynamics of relative contributions to ER and FIs. In Control, the contribution of poor subjects is relatively stable, while the contribution of rich subjects fluctuates in the last few periods. In Treatment, the contribution of poor subjects to ER is consistent with the contribution of rich subjects to FIs.

We then separate contributions by poor and rich subjects during the first half (periods 1–3 for T = 6 and periods 1–5 for T = 10) and second half (the remaining periods in both cases) of the experiments. Results are compiled in Table 1, where we also disentangle the successful groups from the failed groups. In all sessions, the poor never contribute significantly more in the second half, the opposite happening with the rich, except in failed Control groups. In Treatment, the rich contribute to ER (FIs) less (more) during the first half. Furthermore, the rich contribute little to ER in successful groups under Treatment, whereas in failed groups they contribute comparatively more to ER, despite their aggregate contribution (ER + FI) being smaller in failed groups compared with successful groups. The poor, in turn, contribute more to ER in successful groups (in fact, their contribution to ER is determinant in reaching the target); furthermore, their contributions in the second half remain nearly unaffected compared with the first half in successful groups, whereas they decline significantly in failed groups. Overall, these results suggest the importance of having sustained contributions (of the poor to ER and the rich to FIs) during the entire experiment to warrant success.

DISCUSSION

Given the heterogeneity of countries in what concerns both wealth and ER costs, and given the difficulty in implementing effective binding enforcement mechanisms in international agreements, the implementation of an FI, along lines similar to the so-called Financial Mechanism of the UNFCCC, constitutes a possible way out of the global ER conundrum that, up to now, remained unexplored. Our theoretical analysis predicts that cooperation can become stable (albeit with an associated narrow basin of attraction) only in the presence of the FI (incentive NE), provided there is a positive correlation between the contributions of the rich to FIs and the contributions of the poor to ER (Note S4). The results from our behavioral experiments, in addition to unequivocally confirming such a positive correlation, also indicate that humans largely exceed these scanty expectations. Deviating from rational behavior, humans often exhibit a surprising cooperative potential that should not be overlooked.[44] As a result, in the presence of the FI, most mitigation actually stems from the poor, while the rich mostly contribute to FIs. The results further show that the enhancement of FI-induced ER is robust to different monitoring periods, risks of collective failure, and ratios of mitigation cost factors between rich and poor subjects.

That said, it is important to stress under what conditions our results were obtained: the FI mechanism introduced here is

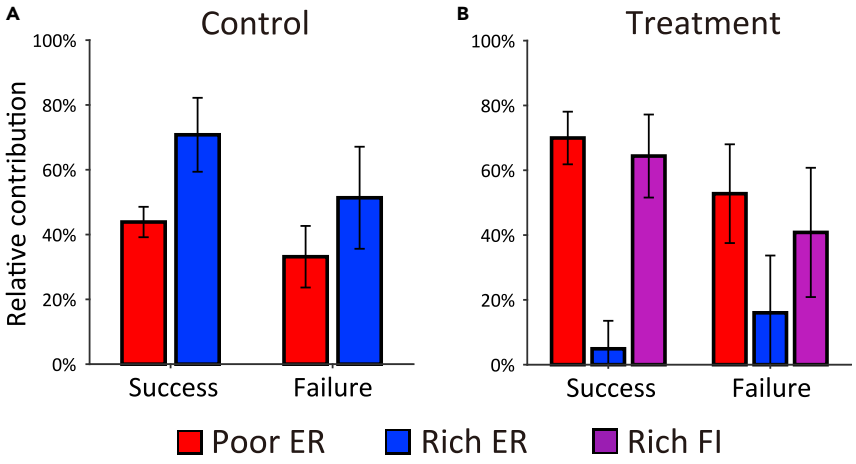


Figure 3. Average relative contributions in successful and failed groups at the group level

Data are presented as mean ± standard deviation. We use the same color codes as in Figure 2.

(A) Aggregated data stemming from all experiments in Control, with separation of successful (reached target) from failed (did not reach it) groups. Both rich and poor subjects invested more in ER in the successful than in the failed groups (Mann-Whitney U test, $p < 0.01$ for rich subjects and $p = 0.01$ for poor subjects). The average relative contributions for poor and rich subjects in successful and failed groups are 0.439 and 0.708, and 0.332 and 0.514, respectively, and the standard deviations are 0.047 and 0.095, and 0.114 and 0.158, respectively.

(B) Aggregated data stemming from all experiments in Treatment, with separation of successful from failed groups. Poor subjects invested more in ER and rich subjects contributed more to FIs in successful groups (Mann-Whitney U test, $p < 0.01$ for both rich and poor subjects), while rich subjects in failed groups invested more in ER (Mann-Whitney U test, $p < 0.01$). The average relative contributions for poor subjects to ER and rich subjects to ER and FIs in successful and failed groups are 0.7, 0.049, and 0.644 and 0.528, 0.16, and 0.408, respectively, and the standard deviations are 0.081, 0.086, and 0.128 and 0.152, 0.177, and 0.2, respectively.

purely endogenous, in the sense that transfers are budget-balanced. Furthermore, poor subjects cannot invest more per period even if their wealth increases via FIs. While this feature allows for a better control of both experiments and theory (as well as for a better comparison between experiments carried out using different parameters), it is easy to anticipate that more flexible forms of FI both are easy to design and will have the potential to warrant better results. Thus, we expect the present results to provide a lower bound, allowing, e.g., poor subjects to use their resources (including those obtained via FIs) at will may promote additional ER. We further note that the incentive provided by rich subjects is assumed to be distributed among poor subjects proportional to their ER contributions. In reality, accurate assessment of ER may be difficult, and hence, the distribution may not be perfectly proportional. Theoretical analysis predicts that the incentive equilibrium exists provided the (expected) amount of incentive to contributing poor subjects is larger than a threshold value (computed explicitly in Note S3). Thus, we expect FIs to remain effective provided they are sufficient, even when small deviations from proportionality occur.

Overall, FIs lead not only to higher ER, but also to higher group payoff (Table 2). The group average payoff (here associated with wealth) increases 66% from Control to Treatment. In particular, the average payoff of poor subjects in Treatment is more than twice of that in Control (increases 108%), whereas the payoff for rich subjects in Treatment is slightly higher than Control (increases 12%). Given that the FI implemented here is budget-balanced (it does not involve any external funding), this indeed constitutes a Pareto improvement of social welfare regarding the ER problem.

Comparisons between the 6 and 10 period cases (Tables S1 and S2) suggest that 5 years separating monitoring periods resulting from the global stocktake of the PA may be almost ideal, as one expects approximately 6 monitoring periods to take place from 2023 to 2050. Indeed, larger sequences of monitoring events within the same framework (as considered here in the 10 period experiments, simulating, e.g., a possible rescheduling from 2050 to a future date or, alternatively, shorter time intervals between monitoring periods) may prove to be redundant, in the sense that no significant improvement in ER may take place (in absolute terms). In addition, increasing the risks of both rich and poor subjects can lead to more effective ER. Specifically, in Control, rich subjects contribute more to ER (not significant for poor subjects), and in Treatment, rich subjects contribute more to FIs and poor subjects contribute more to ER (Tables S1 and S3).

It is worth noting that the system remains fragile under FI. In 62% of successful Treatment groups (26 of 42), the average ER is marginally (5%) higher than the pre-defined target. Interestingly, in 33% of failed Treatment groups (6 of 18), the total ER achieved is also marginally smaller (5%) than the target. Given the positive correlation found between contributions to FI and ER, it is desirable that rich countries effectively adhere to FIs for this mechanism to work. Simultaneously, if poor countries keep investing in ER during the whole duration of the agreement, FIs may transform into a robust mitigation mechanism. That said, it is important to keep in mind that (1) our experiments took place in a scenario where the threshold to be surpassed was well defined; however, threshold uncertainty, if large, has been shown to change both the nature of the game and the players' behavior, making it more difficult to attain success.[29,45] (2) On the other hand, the fact that our experimental design does not allow players to make pledges creates harsher conditions compared with what is known already[28] and effectively implemented. (3) Finally, in our experiments, there is only one rich subject in a group. Additional experiments showed that having more than one rich player in the group will contribute to "diluting" each one's responsibility, rendering coordination toward the goal more difficult (Figure S5 and Table S4). Notwithstanding, even in these experiments FIs systematically improve both total ER and the rates of success.

Needless to say, our behavioral experiments involving groups of six subjects may not represent the real players of the climate game. Specifically, achieving cooperation is more difficult for larger than for smaller groups,[32,40,46] as well as when more than one high-income country participates in a group (see

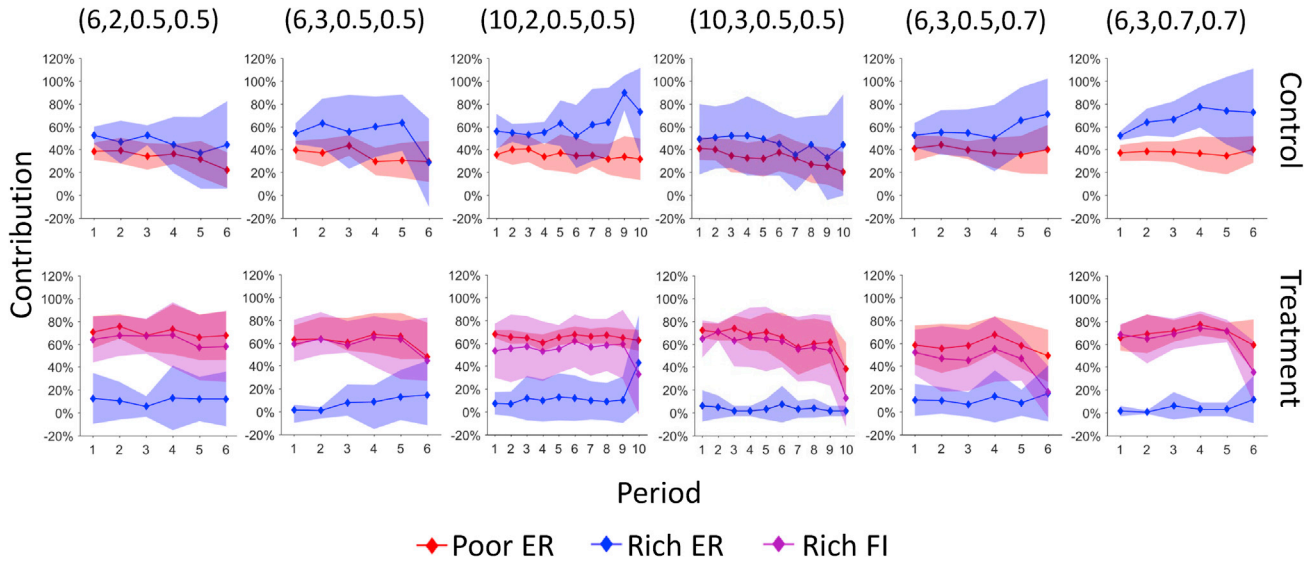


Figure 4. Dynamics of average relative contributions to ER and FIs in experiments with different T, s, r_R, and r_P at the group level

Data are presented as mean ± standard deviation. We use the same color conventions of Figure 2. Top: Control. Bottom: Treatment. The average values (solid diamonds) and standard deviations (shaded areas) of relative contributions were computed at the group level. See Tables S6 and S7 for details of the average relative contributions and standard deviations of 6-period experiments and 10-period experiments.

Note S6 for additional experiments). Indeed, in the latter case, the possibility of free riding among the high-income representatives potentially shadows the success obtained when only one high-income country is present in the group. Our results stress the importance of witnessing coordinated action among the high-income participants in climate negotiations, as they play a pivotal role in determining overall success. These examples suggest that our findings may underrate the severity of the ER dilemma. Thus, even if all groups reach the target, we cannot conclude that FIs will succeed. Conversely, if most groups fail to reach the target, then it would be more difficult to achieve the ER target in reality. However, the fact that our experiments involve small groups, similar to most behavioral experiments on climate change carried out to date,[24,28,40,47] does not mean that the results obtained lack important insights. Indeed, analyzing the individual behaviors in successful and failed groups provides important guidelines for how to implement FIs efficiently. On one hand, both experiment and theory indicate that the FI works only if there is a positive correlation between contributions to the FI by the rich and mitigation by the poor. Thus, a policy implication is that, in practice, developed countries and developing countries should carefully negotiate and, it is hoped, improve the design of the FI to allow for an efficient support of ER, in line with the principle of "common but differentiated responsibilities." Last, but not least, the asymmetric behaviors between the rich and the poor under FIs may act to change the norms at work in global mitigation,[19,38] also inhibiting potential homophilic behavior among the rich and the poor, a feature that was found to play a crucial role in promoting cooperation toward global ER.[32]

We would like to point out that the present model, similar to others, contains some simplifying assumptions that constitute an oversimplification of the real-world case. That said, some of the insights here provide may prove useful in designing FIs in connection with the climate change problem (and general multi-period threshold public goods games). In the future we plan to investigate the effectiveness of FIs in situations in which poor subjects are allowed to adopt heterogeneous strategies; the game may have an uncertain number of periods, in which contributions by one group may be observed by another group; and groups may have different sizes and/or different compositions.

EXPERIMENTAL PROCEDURES

Resource availability

We are willing to distribute materials and protocols to qualified researchers, details are as follows.

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Boyu Zhang (zhangby@bnu.edu.cn).

Materials availability

Materials generated in this study will be made available on reasonable request.

Data and code availability

Raw data of the behavioral experiments (Controls 1–6, Treatments 1–6, Control 2R, Treatment 2R) can be downloaded from Mendeley Data: https://doi.org/10.17632/bf8zmpmj4w.1 https://data.mendeley.com/datasets/bf8zmpmj4w/1.

Procedures and model

Here we summarize the experimental procedures as well as the theoretical model; further details, as well as a discussion of groups containing more than a single rich subject, are provided in Notes S1 and S2.

The experimental protocols adhered to the standards set by the Declaration of Helsinki and were approved by the local research ethics committee at the State Key Laboratory of Cognitive Neuroscience and Learning, Beijing Normal University, Beijing, China, with reference number CNL_A_0007_001. All participants provided written informed consent to participate after the experimental procedures had been fully explained and acknowledged their right to withdraw at any time during the experiment.

In line with previous experiments,[23,24,28,29,33,38,42] we conducted 12 experimental sessions, including 6 Control sessions and 6 Treatment sessions, where each session consisted of 10 groups of 6 subjects. Group members,

Table 1. Average contributions in the first half (periods 1–3 for T = 6 and periods 1–5 for T = 10) and second half (periods 4–6 for T = 6 and periods 6–10 for T = 10) of the experiments at the group level

First half/second half	Control successful	Control failed	Treatment successful	Treatment failed
Poor ER	1.75/1.76*	1.47*/1.18	2.85*/2.81	2.14*/1.86
Rich ER	12.68/15.64*	10.52*/10.02	0.89/1.22*	2.15/4.16*
Rich FI			13.73*/12.25	8.19*/7.07

Larger value in each entry is highlighted by an asterisk. In failed groups, poor subjects contributed less in the second half (Mann-Whitney U test, p = 0.02 in Control and p = 0.48 in Treatment).

each acting as a delegate representing one country, engaged in what we term the global greenhouse-gas ER tournament. In each group, 1 subject was randomly assigned the rich country role and the remaining 5 assumed poor country roles; roles did not change during each session. In addition, we carried 2 sessions, where each session involved 5 groups of 12 subjects (Control 2R and Treatment 2R) and each group had 2 rich subjects and 10 poor subjects (Note S6). We took these group compositions because there are about 40 Annex I countries (developed countries) and about 200 countries in total in the PA, i.e., the ratios of Annex I and Non-Annex I countries are about 0.2 and 0.8, respectively.[48] In accordance with the agreement, there are $6 \times 0.2 \approx 1$ rich subject, and $6 \times 0.8 \approx 5$ poor subjects in the 6 subject group, and $12 \times 0.2 \approx 2$ rich subjects, and $12 \times 0.8 \approx 10$ poor subjects in the 12 subject group.

Below we characterize endowments and investments in terms of points, where the actual conversion was 1 point = 1 Chinese yuan. Every participant walked away with an amount of Chinese yuan equal in number to the amount of points left at the end of the experiment plus 30. Further details are provided in Note S1.

In Control, there is no FI. A rich subject can contribute from 0 to 20 points (integer values only) to ER in each period—with a cost factor s = 2 or s = 3—and a poor subject can contribute 0 to 4 points (with cost factor s = 1). The total endowments for a rich subject are 120 and 200 points in games with periods T = 6 and T = 10, respectively, while each poor subject gets 24 and 40, respectively.

In Treatment, each of the T periods comprises an additional FI stage before the ER stage, in which a rich subject can contribute (points) to FIs that will be later distributed by the poor subjects, and the poor subjects know the contribution to the FI when they make decisions in the ER stage (see screen shots in Figure S6 for details). No more than a total of 20 points can be used by each rich individual, per period, in the two stages, FI and ER. The contribution to FIs made by the rich will be distributed among the five poor subjects in proportion to their contributions in the subsequent ER stage (rounded up to one decimal). Specifically, if all poor subjects contribute 0, then the they will share the contribution equally. Following previous experiments on public goods games, poor subjects cannot invest more than 4 points per period even if their wealth increases via FI.[35,42] To achieve the ER target, six group members need to contribute half of their initial endowment to ER (which implies fixing the targets in games with (T, s) = (6, 2), (6, 3), (10, 2), and (10, 3) to 90, 80, 150, and 133, respectively). If the total ER target is attained at the end of the game, then subjects' final scores are their remaining points. If not, rich subjects lose all their savings with probability r_R and poor subjects lose their savings with probability r_P (r_R and r_P are taken as 0.5 or 0.7). We note that the ER targets for different parameter settings are different, but are the same for Control and Treatment under a certain parameter setting (see Table S5 for details of experimental settings).

Table 2. Payoffs of subject types in different types of experiments

Type	Poor payoff	Rich payoff	Group payoff	Total ER
Control	38%	25%	32%	82%
Treatment	79%	28%	53%	98%

Values tabulated are average values (aggregating data from Controls and Treatments 1–6), represented as the fraction of the initial endowments of rich and poor subjects, except the last column, where percentages represent percentual values relative to the target ER values in each game. FIs not only increase the total ER, but also increase the payoffs of both rich and poor subjects.

In each period of a session, information gathered in the previous period is given, including contributions by group members and the subject's own score. Note that the cumulative sum of contributions was not displayed on a computer screen. Instead, the subjects were given pen and paper, and they were encouraged to take notes during the game.[24] Sample instructions and further details can be found in Note S1, and screen shots of experimental interfaces are shown in Figure S6.

Theoretical model

In the following, we discuss a model that closely mimics the experiments performed, involving T periods and groups with six subjects, one of which is rich, and the remaining five are poor subjects. In each period, a rich subject can contribute from 0 to 20 points to ER (with an abatement cost factor $s > 1$), and a poor subject can contribute 0 to 4 points.

In Control, we denote a poor subject j's strategy by $\mathbf{C}_{Pj} = (C_{Pj,1}, \ldots, C_{Pj,T}, D_{Pj,1}, D_{Pj,T-1})$ with $j = 1, 2, 3, 4, 5$ and $C_{Pj,t} = 0, 1, 2, 3, 4$, where $C_{Pj,t}$ is the contribution in period t if the total ER in the previous $t-1$ periods in not less than $D_{Pj,t-1}$ (if the total ER is less than $D_{i,t-1}$, then subject j will not contribute in period t and all later periods), and the rich subject's strategy by $\mathbf{C}_R = (C_{R,1}, \ldots, C_{R,T}, D_{R,1}, D_{R,T-1})$ with $C_{R,t} = 0, 1, \ldots, 20$ (the subscript i of $C_{Ri,t}$ and $D_{Ri,t}$ is omitted, as there is only one rich subject), where $C_{R,t}$ is the contribution in period t if the total ER in the previous $t-1$ periods in not less than $D_{R,t-1}$.

Given the multi-period structure of this threshold public goods dilemma, individual strategies can be very complicated.[30] In the following, for the sake of this theoretical analysis, we focus on a specific subset of SPNE that we designate as QSPNE, where all poor countries use the same strategy (i.e., $C_{Pj,t} = C_{P,t}$ for all $j = 1, 2, 3, 4, 5$). In Note S3, we show that strategies of poor and rich subjects at a QSPNE can be characterized by their total contributions $\widehat{C}_P = \sum_{t=1}^{T} C_{P,t}$ and $\widehat{C}_R = \sum_{t=1}^{T} C_{R,t}$, respectively. Furthermore, theorem S1 in Note S3 indicates that Control entails two classes of SPNE: the defective SPNE $(\widehat{C}_P^0, \widehat{C}_R^0) = (0, 0)$ and a set of cooperative QSPNE $(\widehat{C}_P^*, \widehat{C}_R^*) = (2T, 10T)$. Naturally, the ER target is achieved only at the cooperative QSPNE.

In Treatment, a rich subject's strategy is denoted as $(C_{R,1}, \ldots, C_{R,T}, I_1, \ldots, I_T, D_{R,1}, D_{R,T-1})$, where I_t is the contribution to the FI in period t. In addition, we assume that no more contributions to the FI take place when the group reaches the target, because in this case no further ER is needed (this, as is well known, constitutes a "rational" assumption that is often violated in practice). In contrast, a poor subject j's strategy is denoted as $(C_{Pj,1}, \ldots, C_{Pj,T}, D_{Pj,1}, D_{Pj,T-1}, L_{Pj,1}, \ldots, L_{Pj,T})$. Like $D_{Pj,t}$, $L_{Pj,t}$ is also a contribution threshold, where subject j contributes $C_{Pj,t}$ in period t only if the total ER in the previous $t-1$ periods in not less than $D_{Pj,t-1}$ and the total amount contributed in the FI stage in period t, I_t, is not less than $L_{Pj,t}$. If $\sum_{k=1}^{t}(C_{R,k}/s + \sum_{j=1}^{5} C_{Pj,k}) < D_{Pj,t}$ or $I_t < L_{Pj,t}$, then the poor subject will contribute 0 in period t and all later periods. Following this notation, poor subject j receives $I_t C_{P,jt} / \sum_{k=1}^{5} C_{Pk,t}$ in period t. Thus, a defective poor subject who contributes 0 does not receive anything (except that all poor subjects contribute 0). This setting establishes a positive correlation between contribution to FIs by the rich and mitigation by the poor, where more contributions to FIs could lead to more mitigation.

The defective SPNE and the cooperative QSPNE in Control remain SPNE in Treatment. In addition to these two classes of equilibria, Treatment also exhibits a new class of cooperative QSPNE that we designate as incentive QSPNE, where only poor subjects contribute to ER and the rich subject incentivizes the poor subjects by contributing to the FI. In Note S3, we show that the

strategies of poor and rich subjects at a QSPNE in Treatment can be characterized by $\widehat{C}_P = \sum_{t=1}^{T} C_{P,t}$, $\widehat{C}_R = \sum_{t=1}^{T} C_{R,t}$, and $\widehat{I} = \sum_{t=1}^{T} I_t$, where $\widehat{I}$ is the total contributions of the rich subject to the FI. Furthermore, Theorem S2 in Note S3 indicates that at an incentive QSPNE, the strategies for poor and rich subjects can be denoted by $\widehat{C}_P^{**} = 2(1/s+1)T$, $\widehat{C}_R^{**} = 0$, and $\widehat{I}^{**}$ with $10(1-2r_P+1/s)T \leq \widehat{I}^{**} \leq 20r_RT$.

Since these games have multiple equilibria, it is important to investigate which equilibria are more likely to be selected. To answer this question, we carry out a simplified analysis of the evolutionary stability of these equilibria by using replicator dynamics assuming infinite well-mixed populations (Note S4). In Control, the selfish non-cooperative SPNE always dominate the cooperative QSPNE for $r_R = r_P = 0.5$(Figure S1). In Treatment, however, the incentive QSPNE may become locally stable whenever $10(1-2r_P+1/s)T \leq \widehat{I}^{**} \leq 20r_RT$ with an associated basin of attraction, which is always much smaller than that associated with the defective SPNE (Figure S2). Furthermore, higher contributions to FI by the rich promote higher investments in ER by the poor.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.oneear.2021.07.006.

ACKNOWLEDGMENTS

We acknowledge the financial support of the National Natural Science Foundation of China (grants 72091511, 11975049, 71631002, 71771026, and 71922004), The Fundamental Research Funds for the Central Universities of China, and Fundação para a Ciência e a Tecnologia (FCT) Portugal (grants PTDC/MAT-APL/6804/2020, PTDC/CCI-INF/7366/2020, and UIDB/04050/2020). We thank Yafei Wang, Bin Chen, Chunlei Yang, Yi Tao, Sai Liang, and Xuefeng Cui for helpful discussions.

AUTHOR CONTRIBUTIONS

Conceptualization, Y.D., B.Z., W.W., and J.M.P.; methodology, Y.D., B.Z., W.W., and J.M.P.; software, S.M.; formal analysis, Y.D., B.Z., W.W., and J.M.P.; investigation, Y.D., S.M., B.Z., and W.W.; data curation, Y.D., B.Z., and W.W.; writing – original draft, Y.D., B.Z., W.W., and J.M.P.; writing – review & editing, Y.D., B.Z., W.W., and J.M.P.; supervision, B.Z., W.W., and J.M.P.; funding acquisition, B.Z. and W.W.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: December 14, 2020
Revised: May 31, 2021
Accepted: July 19, 2021
Published: August 10, 2021

REFERENCES

1. UNFCCC (2015). Adoption of the Paris agreement. Report No. FCCC/CP/2015/L.9/Rev.1. http://unfccc.int/resource/docs/2015/cop21/eng/l09r01.pdf.
2. Falkner, R. (2016). The Paris Agreement and the new logic of international climate politics. Int. Aff. *92*, 1107–1125.
3. Hale, T. (2016). "All hands on deck": the Paris Agreement and nonstate climate action. Glob. Environ. Polit. *16*, 12–22.
4. Klein, D., Carazo, M.P., Doelle, M., Bulmer, J., and Higham, A. (2017). The Paris Agreement on Climate Change: Analysis and Commentary (Oxford University Press).
5. UNFCCC (2010). Cancun agreements. 16th Conference of the Parties. Cancun: United Nations. http://unfccc.int/meetings/cancun_nov_2010/meeting/6266.php.
6. Christoff, P. (2016). The promissory note: COP 21 and the Paris climate agreement. Environ. Polit. *25*, 765–787.
7. IPCC (2007). Climate Change 2007: Mitigation of Climate Change, Summary for Policymakers. Contribution of Working Group III to the Fourth Assessment (Cambridge University Press).
8. World Bank. https://data.worldbank.org/indicator/NY.GDP.MKTP.CD.
9. World Bank. https://data.worldbank.org/indicator/EN.ATM.CO2E.KD.GD.
10. IPCC (2012). Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation Special Report of the Intergovernmental Panel on Climate Change (Cambridge University Press).
11. Lyster, R. (2017). Climate justice, adaptation and the Paris Agreement: a recipe for disasters? Environ. Polit. *26*, 438–458.
12. Rogelj, J., Den Elzen, M., Höhne, N., Fransen, T., Fekete, H., Winkler, H., and Meinshausen, M. (2016). Paris Agreement climate proposals need a boost to keep warming well below 2 °C. Nature *534*, 631–639.
13. Ostrom, E. (1990). Governing the Commons: The Evolution of Institutions for Collective Action (Cambridge University Press).
14. Carraro, C., and Siniscalco, D. (1993). Strategies for the international protection of the environment. J. Public Econ. *52*, 309–328.
15. Nordhaus, W.D. (1994). Managing the Global Commons: The Economics of Climate Change (MIT Press).
16. Carraro, C. (2007). Incentives and institutions: a bottom-up approach to climate policy. In Architectures for Agreement: Addressing Global Climate Change in the Post-Kyoto World, J. Aldy and R.N. Stavins, eds. (Cambridge University Press), pp. 161–172.
17. Ostrom, E. (2014). A polycentric approach for coping with climate change. Ann. Econ. Finance *15*, 97–134.
18. Barrett, S. (2011). Avoiding disastrous climate change is possible but not inevitable. Proc. Natl. Acad. Sci. U S A *108*, 11733–11734.
19. Stewart, R.B., Oppenheimer, M., and Rudyk, B. (2013). A new strategy for global climate protection. Climat. Chang. *120*, 1–12.
20. Nordhaus, W. (2015). Climate clubs: Overcoming free-riding in international climate policy. Am. Econ. Rev. *105*, 1339–1370.
21. Nyborg, K. (2016). Social norms as solutions. Science *354*, 42–43.
22. Hannam, P.M., Vasconcelos, V.V., Levin, S.A., and Pacheco, J.M. (2017). Incomplete cooperation and co-benefits: Deepening climate cooperation with a proliferation of small agreements. Climat. Chang. *144*, 65–79.
23. Milinski, M., Semmann, D., Krambeck, H.J., and Marotzke, J. (2006). Stabilizing the Earth's climate is not a losing game: Supporting evidence from public goods experiments. Proc. Natl. Acad. Sci. U S A *103*, 3994–3998.
24. Milinski, M., Sommerfeld, R.D., Krambeck, H.J., Reed, F.A., and Marotzke, J. (2008). The collective-risk social dilemma and the prevention of simulated dangerous climate change. Proc. Natl. Acad. Sci. U S A *105*, 2291–2294.
25. Pacheco, J.M., Santos, F.C., Souza, M.O., and Skyrms, B. (2009). Evolutionary dynamics of collective action in n-person stag hunt dilemmas. Proc. R. Soc. B *276*, 315–321.
26. Milinski, M., Röhl, T., and Marotzke, J. (2011). Cooperative interaction of rich and poor can be catalyzed by intermediate climate targets. Climatic Change *109*, 807–814.
27. Santos, F.C., and Pacheco, J.M. (2011). Risk of collective failure provides an escape from the tragedy of the commons. Proc. Natl. Acad. Sci. U S A *108*, 10421–10425.
28. Tavoni, A., Dannenberg, A., Kallis, G., and Löschel, A. (2011). Inequality, communication and the avoidance of disastrous climate change in a public goods game. Proc. Natl. Acad. Sci. U S A *108*, 11825–11829.
29. Barrett, S., and Dannenberg, A. (2012). Climate negotiations under scientific uncertainty. Proc. Natl. Acad. Sci. U S A *109*, 17372–17376.
30. Chakra, M.A., and Traulsen, A. (2012). Evolutionary dynamics of strategic behaviour in a collective-risk dilemma. PLoS Comput. Biol. *8*, e1002652.
31. Vasconcelos, V.V., Santos, F.C., and Pacheco, J.M. (2013). A bottom-up institutional approach to cooperative governance of risky commons. Nat. Clim. Change *3*, 797–801.

32. Vasconcelos, V.V., Santos, F.C., Pacheco, J.M., and Levin, S.A. (2014). Climate policies under wealth inequality. Proc. Natl. Acad. Sci. U S A *111*, 2212–2216.

33. Barrett, S., and Dannenberg, A. (2014). Sensitivity of collective action to uncertainty about climate tipping points. Nat. Clim. Change *4*, 36–39.

34. Milinski, M., Hilbe, C., Semmann, D., Sommerfeld, R., and Marotzke, J. (2016). Humans choose representatives who enforce cooperation in social dilemmas through extortion. Nat. Commun. *7*, 10915.

35. Perc, M., Jordan, J.J., Rand, D.G., Wang, Z., Boccaletti, S., and Szolnoki, A. (2017). Statistical physics of human cooperation. Phys. Rep. *687*, 1–51.

36. Andrews, T.M., Delton, A.W., and Kline, R. (2018). High-risk high-reward investments to mitigate climate change. Nat. Clim. Change *8*, 890.

37. Chakra, M.A., Bumann, S., Schenk, H., Oschlies, A., and Traulsen, A. (2018). Immediate action is the best strategy when facing uncertain climate change. Nat. Commun. *9*, 2566.

38. Kline, R., Seltzer, N., Lukinova, E., and Bynum, A. (2018). Differentiated responsibilities and prosocial behaviour in climate change mitigation. Nat. Hum. Behav. *2*, 653–661.

39. Gross, J., and De Dreu, C.K. (2019). Individual solutions to shared problems create a modern tragedy of the commons. Sci. Adv. *5*, eaau7296.

40. Wang, Z., Jusup, M., Guo, H., Shi, L., Geček, S., Anand, M., Perc, M., Bauch, C.T., Kurths, J., Boccaletti, S., and Schellnhuber, H.J. (2020). Communicating sentiment and outlook reverses inaction against collective risks. Proc. Natl. Acad. Sci. U S A *117*, 17650–17655.

41. Nordhaus, W. (2013). The Climate Casino: Risk, Uncertainty, and Economics for a Warming World (Yale University Press).

42. Rand, D.G., Dreber, A., Ellingsen, T., Fudenberg, D., and Nowak, M.A. (2009). Positive interactions promote public cooperation. Science *325*, 1272–1275.

43. Hofbauer, J., and Sigmund, K. (1998). Evolutionary Game and Population Dynamics (Cambridge University Press).

44. Simon, H.A. (1982). Models of Bounded Rationality: Empirically Grounded Economic Reason (MIT Press).

45. Pacheco, J.M., Vasconcelos, V.V., and Santos, F.C. (2014). Climate change governance, cooperation and self-organization. Phys. Life Rev. *11*, 573–586.

46. Boyd, R., and Richerson, P.J. (1988). The evolution of reciprocity in sizable groups. J. Theor. Biol. *132*, 337–356.

47. Jacquet, J., Hagel, K., Hauert, C., Marotzke, J., Röhl, T., and Milinski, M. (2013). Intra-and intergenerational discounting in the climate game. Nat. Clim. Change *3*, 1025–1028.

48. (2013). UNFCCC. https://unfccc.int/process/parties-non-party-stakeholders/parties-convention-and-observer-states.

One Earth, Volume *4*

Supplemental information

Financial incentives to poor countries promote net emissions reductions in multilateral climate agreements

Yali Dong, Shuangmei Ma, Boyu Zhang, Wen-Xu Wang, and Jorge M. Pacheco

Supplemental Information

Supplemental Note 1. Experimental design

1.1. Details of experimental settings

We conducted 12 sessions of experiments, where each session involved 10 different groups of 6 subjects. In the 6 Control sessions, only an **ER** stage takes place in each period with no **FI** stage involved; in the 6 Treatment sessions, there is an **FI** stage before the **ER** stage in each period. Detailed settings of the experiments are provided in SI Table S5. In addition, we carried 2 sessions, each involving 5 groups of 12 subjects (Control 2R and Treatment 2R). In the Control 2R and Treatment 2R sessions, each group has 2 rich subjects and 10 poor subjects (see SI Note 6 below for details).

Experiments were conducted in computer labs of Beijing Normal University. All 840 subjects were undergraduates and graduates recruited from Beijing Normal University who had not taken classes on game theory and economics. The interactions were anonymous, and took place via computers. Frosted glass dividers ensured that the students could not see each other. The experimental platform software was locally built using PHP, mySQL and javascript, and ran locally on the servers. Schematic diagrams of our experimental platform are shown in SI Figure S6.

Before starting each experiment, we used 20 minutes to explain the game to all subjects, including the climate change problem and the rules of the computer game. All subjects in each session were given the same instructions (in Chinese). In accord with previous studies, the instructions were framed in the context of climate change and its mitigation. Along with the instructions, a training period was included, two examples are introduced. In the first, poor subjects contribute 2 and rich subjects contribute 10 to **ER** each period. In the second, poor subjects contribute 3 to **ER** and rich subjects contribute 15 to **FI** (to incentivize poor subjects) each period. Subjects were asked to calculate the total **ER** for each case. All subjects are free to ask questions, and our experimenters would answer all their questions. At the end of the training period a few routine questions based on the two examples were asked to the participants to make sure they all understood the rules of game.

There is no time limit for decision making, and each experiment lasted about 20 minutes. After the experiment, the total scores (measured in points, see Methods section) of each subject were converted to Chinese Yuan at a ratio of 1:1. This amount,

plus 30 Chinese Yuan constituted the final income of each participant.

1.2. The FI mechanism and the Paris Agreement

Here we detail the aspects of the Paris Agreement (**PA**) that motivated our experimental design and implementation of the **FI** mechanism and global **ER**.

Up to 2017, nearly 200 countries (and regions) became members of the United Nations Framework Convention on Climate Change (UNFCCC), and signed the **PA**. In our experiments, we follow the idea of Pareto Principle (or the 80-20 rule) to set the proportion of rich subjects to one-sixth, and their total wealth to be equal to the accumulated wealth of poor subjects. Specifically, the initial endowment of a rich subject is 5 times that of a poor subject.

Meanwhile, the cost of **ER** is typically higher for rich (developed) countries than poor (developing) countries. For instance, in 2014, the Gross Domestic Product (**GDP**) per kg of CO2 emissions — used as a surrogate for cost of **ER** — was 3.2 times higher in high-income countries compared to that in middle- and low-income countries. In this regard, we set two high **ER** costs for a rich subject, i.e., s=2 and s=3, respectively, and set a low **ER** cost s=1 for every poor subject.

In the **PA**, countries agreed to pledge their mitigation progresses every 5 years, starting in 2023 (called "Intended Nationally Determined Contributions" and "Global Stocktake", see Articles 4 and 14 in the **PA**). Developed countries could voluntarily provide funds to help undeveloped countries adapt and mitigate (called "Financial Mechanism", see Articles 5 and 9 in the **PA**). In addition, contributions from developed to developing countries will be monitored, such that the future decisions of developed countries may be contingent on perceived status. In the **PA**, the global emissions ought to be reduced to 50% of the present level by 2050. In other words, after approximately 6 monitoring periods, the target should be achieved. Thus, we consider two values for the number of monitoring periods, T=6 as in the **PA** and T=10 for a longer process (or shorter time intervals between consecutive monitoring periods).

Unlike Intended Nationally Determined Contributions and Global Stocktake, the **PA** only gives a few implementation details on the Financial Mechanism. Below, we list the key aspects that were incorporated in the treatment experiments.

Key aspect 1: Positive incentives are encouraged to support developing countries based on their emission reduction results.

Paragraph 2 in Article 5: "Parties are encouraged to take action to implement and

support, including through results-based payments, the existing framework as set out in related guidance and decisions already agreed under the Convention for: policy approaches and positive incentives for activities relating to reducing emissions from deforestation and forest degradation, and the role of conservation, sustainable management of forests and enhancement of forest carbon stocks in developing countries;"

Key aspect 2: Developed countries should provide the incentive.

Paragraph 1 in Article 9: "Developed country Parties shall provide financial resources to assist developing country Parties with respect to both mitigation and adaptation in continuation of their existing obligations under the Convention."

Key aspect 3: The incentive provided by developed countries is also a part of Global Stocktake that can be seen by all other members.

Paragraph 6 in Article 9: "The global stocktake referred to in Article 14 shall take into account the relevant information provided by developed country Parties and/or Agreement bodies on efforts related to climate finance."

Paragraph 7 in Article 9: "Developed country Parties shall provide transparent and consistent information on support for developing country Parties."

Key aspect 4: Reason for the designation "financial mechanism".

Paragraph 8 in Article 9: "The Financial Mechanism of the Convention, including its operating entities, shall serve as the financial mechanism of this Agreement."

In summary, the **FI** included in our Treatment experiments is based on the principles and mechanisms established in the **PA**, as shown above. In addition, two additional assumptions are included in our experiments: (1) the incentive will be distributed among poor subjects in proportion to their contributions to **ER**; (2) the incentive received from rich subjects cannot be used for **ER**.

Assumption (1) seems reasonable provided the amount of **ER** can be monitored; moreover, the distribution rule according to **ER** is promising and fair. Assumption (2) is motivated by the fact that the game framework that we setup is budget balanced. This assumption naturally imposes a harder constraint on the capability of developing countries to mitigate. In other words, removing assumption (2) is likely to increase the

overall capability of countries to solve the climate change problem, insofar as the funds are used exclusively for mitigation.

1.3. Sample instruction

In the following, we shall use the Treatment 1 version of the experiment as the experimental template.

General notice

Welcome and thanks for participating in this game. Please read the game instructions carefully. If you have any question please raise your hand. One experimenter will then come to you and answer your questions. From now on, communication with other participants is not allowed. Please switch off your mobile phones and keep quiet during the experiment. You will play a decision-making game. In the game, you do not know the other persons' true identity. Your scores will depend on you and your partners' decisions. In the end, your income will be calculated in the following way:

Your final income = fixed income 30 Chinese Yuan + your total score.

Climate change

This game intends to simulate the mitigation of greenhouse gas emissions. Global warming is seen as a serious environmental problem faced by mankind. The recent Paris Agreement is a landmark agreement dealing with mitigation of greenhouse gas emissions. It aims to respond to the global climate change threat by keeping an average global temperature rise this century below 2°C above pre-industrial levels and to pursue efforts to limit this temperature increase by no more than 1.5°C. As of October 2017, 195 members of the United Nations Framework Convention on Climate Change (UNFCCC) have signed the agreement.

In the Paris Agreement, each country voluntarily determines its own contribution to mitigate global warming. The contributions should be reported and evaluated every 5 years, but there is no mechanism to force a country to set a specific target by a specific date.

Rules of the game Treatment 1 (T,s,r_R,r_P)=(6,2,0.5,0.5)

6 subjects are involved in a group, and play a game for 6 periods. In each group,

1 subject is randomly assigned to play the role of a rich country and the other 5 will play the role of poor countries. Assigned subject roles do not change during the experiment. A rich subject has 120 initial points and can use 0 to 20 points each period. A poor subject has 24 initial points and can use 0 to 4 points each period.

Each period has two stages: The **FI** stage and the **ER** stage. In the **FI** stage, a rich subject can contribute $I \in [0,20]$ points to a climate fund that will be used to incentivize the poor subjects. This contribution will be distributed among the 5 poor subjects proportionally to their contributions in the next **ER** stage. In the **ER** stage, a rich subject can contribute $C_R \in [0,20 - I]$ points to reduce emission with a cost factor $s = 2$, and a poor subject can contribute $C_P \in [0,4]$ points to reduce emission without discount ($s = 1$). Important: In each period a rich subject can use at most 20 points in both stages (**FI** and **ER**).

If the total emission reduction achieved by the group after 6 periods reaches 90, then your final scores are your left points. If not, then with probability 50% your final scores are your left points; and with probability 50% your final scores are 0.

Supplemental Note 2. Theoretical model

In the following we setup a general theoretical model that we shall use to analyze the experiments just described. They constitute a particular case of the model introduced below.

2.1. Control

Control is an N-person multi-period Threshold Public Goods Dilemma with two types of subjects in a group of size N, where N_1 of them represent rich countries and N_2 of them represent poor countries (with $N_1 + N_2 = N$). The initial endowments for a rich subject and a poor subject are E_R and E_P, respectively, with $E_R > E_P$. In each of T periods, a rich subject can contribute at most E_R/T to **ER**, and a poor subject can contribute at most E_P/T to **ER**. Denoting the contribution of the poor subject j in period t by $C_{Pj,t}$ ($j = 1, \dots, N_2$, $t = 1, \dots, T$, $0 \le C_{Pj,t} \le E_P/T$), then the corresponding **ER** is $C_{Pj,t}$. Let us denote the contribution of the rich subject i in period t by $C_{Ri,t}$ ($i = 1, \dots, N_1$, $t = 1, \dots, T$, $0 \le C_{Ri,t} \le E_R/T$). Because the abatement cost for a rich country is often higher than that for a poor country, the resulting **ER** is $C_{Ri,t}/s$, where s>1 is the abatement cost factor. At the end of the T periods, the total **ER** of the N countries is

$E = \sum_{i=1}^{N_1} \sum_{t=1}^{T} C_{Ri,t}/s + \sum_{j=1}^{N_2} \sum_{t=1}^{T} C_{Pj,t}$.

We denote the target for the **ER** by D, and assume $D > \max\{E_R/s, E_P\}$, i.e., a single country cannot attain the target even if it contributes all its endowments. At the end of T periods, if the total **ER** reaches the target (i.e.,$E \geq D$), then all subjects will retain whatever they did not contribute. Thus, the payoffs for a rich subject and a poor subject are $f_{Ri} = E_R - \sum_{t=1}^{T} C_{Ri,t}$ $(i = 1, \dots, N_1)$ and$f_{Pj} = E_P - \sum_{t=1}^{T} C_{Pj,t}$ $(j = 1, \dots, N_2)$, respectively. If the group failed to reach the target, then rich subjects lose all their savings with probability r_R and poor subjects lose their savings with probability r_P (i.e., r_R and r_P are the risks of failure). Consequently, the expected payoffs for a rich subject and a poor subject are $f_{Ri} = (1 - r_R)(E_R - \sum_{t=1}^{T} C_{Ri,t}$ $(i = 1, \dots, N_1)$ and $f_{Pj} = (1 - r_P)(E_P - \sum_{t=1}^{T} C_{Pj,t})$ $(j = 1, \dots, N_2)$, respectively.

2.2. Treatment

In Treatment, rich subjects can indirectly transfer their endowments to poor subjects, in addition to what was already in place in the Control sessions. To be precise, in each of T periods, there is an **FI** stage before the **ER** stage, in which rich subjects can contribute points to **FI** that will be later distributed by the poor subjects. Denote the contribution of the rich subject i in period t by $I_{i,t}$ $(i = 1, \dots, N_1, t = 1, \dots, T)$. It will be distributed among the poor subjects proportionally to their contributions, i.e., the poor subject j who contributes $C_{Pj,t}$ in period t receives $I_{i,t}C_{Pj,t}/\sum_{k=1}^{N_2} C_{Pk,t}$ from the rich subject i, and receives $\sum_{i=1}^{N_1} I_{i,t}C_{Pj,t}/\sum_{k=1}^{N_2} C_{Pk,t}$ in total in period t. If all poor subjects contribute 0, then they will share the contribution equally. We further assume that $0 \leq I_{i,t} + C_{Ri,t} \leq E_R/T$, i.e., a rich subject can use at most $1/T$ of its initial endowment in each period.

At the end of T periods, if the total **ER** reaches the target, then all countries keep whatever they had not contributed. Thus, the payoffs for a rich subject and a poor subject are

$$f_{Ri} = E_R - \sum_{t=1}^{T}(C_{Ri,t} + I_{i,t}), \quad \text{(S1)}$$

$$f_{Pj} = E_P - \sum_{t=1}^{T}(C_{Pj,t} - \sum_{i=1}^{N_1} I_{i,t}C_{Pj,t}/\sum_{k=1}^{N_2} C_{Pk,t}), \quad \text{(S2)}$$

respectively $(i = 1, \dots, N_1, j = 1, \dots, N_2)$. If the target is not met, then rich subjects lose all their savings with probability r_R and poor subjects lose their savings with probability r_P. Consequently, the expected payoffs for a rich subject and a poor subject

are

$$f_{Ri} = (1 - r_R)(E_R - \sum_{t=1}^{T}(C_{Ri,t} + I_{i,t})), \quad \text{(S3)}$$

$$f_{Pj} = (1 - r_P)(E_P - \sum_{t=1}^{T}(C_{Pj,t} - \sum_{i=1}^{N_1} I_{i,t}C_{Pj,t}/\sum_{k=1}^{N_2} C_{Pk,t})), \quad \text{(S4)}$$

respectively ($i = 1, \dots, N_1$, $j = 1, \dots, N_2$).

Supplemental Note 3. Subgame Perfect Nash equilibrium analysis

Given the multi-period structure of this game, individual strategies can be very complicated and the existence of asymmetric Subgame Perfect Nash Equilibria (**SPNE**) is possible. In the following, instead of attempting to perform a full analysis of the game, we focus on a specific subset of **SPNE** that we designate as Quasi-symmetric **SPNE** (**QSPNE**); at a **QSPNE** all rich subjects use the same strategy and all poor subjects use the same strategy throughout all periods. Furthermore, in the following we shall restrict the **SPNE** analysis to the 6 experimental settings of Control and Treatment (see Figure 1B), that is, we consider a subset of the general game defined above, involving T-periods and groups with 6 subjects, one of which is rich and the remaining 5 are poor subjects. In each period, a rich subject can contribute between 0 to 20 points (always integer values) to **ER** (with an abatement cost factor s>1), and a poor subject can contribute 0 to 4 points (with an abatement cost factor s=1). Thus, the initial endowments for a rich and a poor subject are $E_R = 20T$ and $E_P = 4T$, respectively. The **ER** target is set to $D = 10(1/s + 1)T$. To achieve the **ER** target, each subject needs to contribute on average half of the initial endowment (i.e., the rich contributes 10T and each poor subject contributes 2T). If the total **ER** target is attained at the end, then subjects' final scores are their remaining points. If not, then rich subjects lose their remaining points with probability r_R and poor subjects lose their remaining points with probability r_P.

3.1. Control

Previous studies have identified numbers of **SPNE** strategies that can sustain cooperation under repeated social dilemma game, such as GRIM and Tit-for-Tat (Sigmund, 2010). In this paper, we focus on the sustaining of cooperation in the multi-period Threshold Public Goods Dilemma through GRIM strategy (Abou Chakra, Traulsen, 2012). That is, a subject (denoted by i) using GRIM will contribute to **ER** in period t only if the total **ER** in the previous $t - 1$ periods in not less than a predefined

value $D_{i,t-1}$. If the total **ER** is less than $D_{i,t-1}$, then subject i will not contribute in period t and all later periods.

Following SI Note 2 above, we denote a poor subject j's strategy by $\mathbf{C}_{Pj} = (C_{Pj,1}, \dots, C_{Pj,T}, D_{Pj,1}, D_{Pj,T-1})$ with $j = 1,2,3,4,5$ and $C_{Pj,t} = 0,1,2,3,4$, where $C_{Pj,t}$ is the contribution in period t if the total **ER** in the previous $t-1$ periods in not less than $D_{Pj,t-1}$, and the rich subject's strategy by $\mathbf{C}_R = (C_{R,1}, \dots, C_{R,T}, D_{R,1}, D_{R,T-1})$ with $C_{Rt} = 0,1, \dots, 20$ (the subscript i of $C_{Ri,t}$ and $D_{Ri,t}$ is omitted, as there is only one rich subject), where $C_{R,t}$ is the contribution in period t if the total **ER** in the previous $t-1$ periods in not less than $D_{R,t-1}$. Thus, a strategy profile for Control can be written as $(\mathbf{C}_{P1}, \dots, \mathbf{C}_{P5}, \mathbf{C}_R)$.

Let $f_{Pj}(\mathbf{C}'_{Pj}|\mathbf{C}_{P1}, \dots, \mathbf{C}_{P5}, \mathbf{C}_R)$ denote the payoff for poor subject j if she/he unilaterally changes her/his strategy from $\mathbf{C}_{Pj}$ to $\mathbf{C}'_{Pj}$. Therefore, a strategy profile $(\mathbf{C}_{P1}, \dots, \mathbf{C}_{P5}, \mathbf{C}_R)$ is a (plain) **NE** only if for all $\mathbf{C}'_{Pj} \neq \mathbf{C}_{Pj}$ and $\mathbf{C}'_R \neq \mathbf{C}_R$,

(i) $f_{Pj}(\mathbf{C}'_{Pj}|\mathbf{C}_{P1}, \dots, \mathbf{C}_{P5}, \mathbf{C}_R) \leq f_{Pj}(\mathbf{C}_{P1}, \dots, \mathbf{C}_{P5}, \mathbf{C}_R)$ and

(ii) $f_R(\mathbf{C}_{P1}, \dots, \mathbf{C}_{P5}, \mathbf{C}'_R) \leq f_R(\mathbf{C}_{P1}, \dots, \mathbf{C}_{P5}, \mathbf{C}_R)$.

These two conditions allow us to check whether a strategy is a **NE** for the Control. Furthermore, **SPNE** can be refined from the set of **NE** by the one-shot deviation principle, i.e., a **NE** is also a **SPNE** if all players have no incentive to deviate in any period. Up to this point, poor subjects in the same group might employ different strategies. We now focus on the **QSPNE**, which we denote by $(\mathbf{C}^*_P, \mathbf{C}^*_R)$, such that at all the 5 poor subjects use the same strategy $\mathbf{C}^*_P = (C^*_{P,1}, \dots, C^*_{P,T}, D^*_{P,1}, D^*_{P,T-1})$.

Similar to other social dilemma games, the defective state $(\mathbf{C}^0_P, \mathbf{C}^0_R)$ where $C_{Pj,t} = C_{R,t} = 0$ for all $j = 1,2,3,4,5$ and $t = 1, \dots, T$ is always a **SPNE**. The reason is that any single subject cannot attain the **ER** target by its own effort.

In contrast to the defective **SPNE**, we are more interested in the cooperative **QSPNE**, where both rich and poor subjects contribute. We next provide the **SPNE** conditions for $(\mathbf{C}^*_P, \mathbf{C}^*_R)$ with $C^*_{R,t}, C^*_{P,t} > 0$ for all $t = 1, \dots, T$.

Theorem S1. A strategy profile $(\mathbf{C}^*_P, \mathbf{C}^*_R)$ with $C^*_{R,t}, C^*_{P,t} > 0$ for all $t = 1, \dots, T$ is a **SPNE** for the Control game if (i) $\sum_{t=1}^{T}(C^*_{R,t}/s + 5C^*_{P,t}) = D$, (ii) $D^*_{R,t}, D^*_{P,t} \leq \sum_{k=1}^{t}(C^*_{R,k}/s + 5C^*_{P,k})$ for all $t = 1, \dots, T-1$, (iii) $\sum_{t=1}^{T} C^*_{P,t} \leq r_P E_P$, $\sum_{t=1}^{T} C^*_{R,t} \leq r_R E_R$. At this equilibrium, payoffs for poor and rich subjects are $f_{Pj}(\mathbf{C}^*_P, \mathbf{C}^*_R) = E_P -$

$\sum_{t=1}^{T} C_{P,t}^*$ and $f_R(\mathbf{C}_P^*, \mathbf{C}_R^*) = E_R - \sum_{t=1}^{T} C_{R,t}^*$, respectively.

Proof: We prove the three conditions one by one.

(i) On the one hand, if $\sum_{t=1}^{T}(C_{R,t}^*/s + 5C_{P,t}^*) < D$, i.e., the maximum **ER** is less than the target D, then rich and poor subjects will lose their remaining points with certain probabilities. In this case, a rich (or a poor) subject can get a higher payoff by deviating to the defective strategy $\mathbf{C}_R^0$ (or $\mathbf{C}_P^0$).

On the other hand, if $\sum_{t=1}^{T}(C_{R,t}^*/s + 5C_{P,t}^*) > D$, i.e., the maximum **ER** exceeds the target D, then a rich (or a poor) subject can get a higher payoff by decreasing the contribution in the last period $C_{R,T}^*$ (or $C_{P,T}^*$) (note that decreasing the contribution in the last period does not trigger the GRIM strategy).

Thus, at a cooperative **QSPNE**, we must have $\sum_{t=1}^{T}(C_{R,t}^*/s + 5C_{P,t}^*) = D$.

(ii) If $D_{R,t}^* > \sum_{k=1}^{t}(C_{R,k}^*/s + 5C_{P,k}^*)$ (or $D_{P,t}^* > \sum_{k=1}^{t}(C_{R,k}^*/s + 5C_{P,k}^*)$), then the rich (or the poor) subject will not contribute in period $t-1$ and later periods even if all the players contribute in the previous t periods. From condition (i), the group will fail to reach the target. In this case, a rich (or a poor) subject can get a higher payoff by deviating to the defective strategy $\mathbf{C}_R^0$ (or $\mathbf{C}_P^0$).

(iii) Under conditions (i) and (ii), the total contributions of a poor subject and a rich subject at $(\mathbf{C}_P^*, \mathbf{C}_R^*)$ are $\sum_{t=1}^{T} C_{P,t}^*$ and $\sum_{t=1}^{T} C_{R,t}^*$, respectively. Thus, the payoffs for a poor subject and a rich subject are $f_{Pj}(\mathbf{C}_P^*, \mathbf{C}_R^*) = E_P - \sum_{t=1}^{T} C_{P,t}^*$ and $f_R(\mathbf{C}_P^*, \mathbf{C}_R^*) = E_R - \sum_{t=1}^{T} C_{R,t}^*$, respectively.

Now suppose that a poor subject j deviates from the GRIM in period t. Clearly, this player has no incentive to increase $C_{P,t}^*$ because this cannot lead to a higher payoff. Thus, we only need to consider that this player decreases $C_{P,t}^*$ to 0. In this case, the group fails to meet the target and the maximum expected payoff for subject j is $f_{Pj}(\mathbf{C}_{Pj}'|\mathbf{C}_P^*, \mathbf{C}_R^*) = (1 - r_P)(E_P - \sum_{k=1}^{t-1} C_{P,k}^*)$. Thus, the poor subject j will not deviate in period t if $f_{Pj}(\mathbf{C}_P^*, \mathbf{C}_R^*) \geq f_{Pj}(\mathbf{C}_{Pj}'|\mathbf{C}_P^*, \mathbf{C}_R^*)$, which is equivalent to $r_P E_P \geq r_P \sum_{k=1}^{t-1} C_{P,k}^* + \sum_{k=t}^{T} C_{P,k}^*$. Similarly, the rich subject will not deviate in period t if $r_R E_R \geq r_R \sum_{k=1}^{t-1} C_{R,k}^* + \sum_{k=t}^{T} C_{R,k}^*$. By the one-shot deviation principle, $(\mathbf{C}_P^*, \mathbf{C}_R^*)$ is a **SPNE** if all players have no incentive to deviate in any period. This requires $r_P E_P \geq r_P \sum_{k=1}^{t-1} C_{P,k}^* + \sum_{k=t}^{T} C_{P,k}^*$ and $r_R E_R \geq r_R \sum_{k=1}^{t-1} C_{R,k}^* + \sum_{k=t}^{T} C_{R,k}^*$ for all $t = 1, \ldots, T$. Thus, $(\mathbf{C}_P^*, \mathbf{C}_R^*)$ is a **SPNE** if $\sum_{t=1}^{T} C_{P,t}^* \leq r_P E_P$ and $\sum_{t=1}^{T} C_{R,t}^* \leq r_R E_R$. □

Since the payoffs of poor and rich subjects at a **QSPNE** depends only on their total contribution, we may characterize their strategies by their total contributions $\widehat{C}_{Pj} = \sum_{t=1}^{T} C_{Pj,t}$ and $\widehat{C}_R = \sum_{t=1}^{T} C_{R,t}$, respectively. Theorem S1 implies that at a **QSPNE**, rich and poor subjects contribute at most r_R and r_P of the initial endowment, respectively. Notice that since $\widehat{C}_R^*/s + 5\widehat{C}_P^* = 10(1/s+1)T$ at a cooperative **QSPNE**, the **QSPNE** exists only if $10(1/s+1)T \le r_R E_R/s + 5r_P E_P$. In Control 1, 2, 5, and 6 (i.e., $(T, s, r_R, r_P) = (6,2,0.5,0.5), (6,3,0.5,0.5), (10,2,0.5,0.5)$ and $(10,3,0.5,0.5)$), $r_R E_R/s + 5r_P E_P$ is exactly $10(1/s+1)T$. Therefore, there is a unique class of cooperative **QSPNE** $(\widehat{C}_P^*, \widehat{C}_R^*) = (r_P E_P, r_R E_R) = (2T, 10T)$, where at this **QSPNE** each subject contributes half of its initial endowment. In Control 3 and 4 (i.e., $(T, s, r_R, r_P) = (6,3,0.5,0.7)$ and $(6,3,0.7,0.7)$), $10(1/s+1)T < r_R E_R/s + 5r_P E_P$. Thus, these games exhibit multiple cooperative **QSPNE**, where the classes of **QSPNE** in Control 5 (or Control 6) can be denoted by $(\widehat{C}_P^*, \widehat{C}_R^*)$ with $\widehat{C}_P^* \le r_P E_P = 2.8T, \widehat{C}_R^* \le r_R E_R = 10T$ (or $14T$), and $\widehat{C}_R^*/s + 5\widehat{C}_P^* = 10(1/s+1)T$. In particular, the **QSPNE** in Control 1, 2, 5, and 6 $(\widehat{C}_P^*, \widehat{C}_R^*) = (2T, 10T)$ are also **QSPNE** in Control 3 and 4.

3.2. Treatment

We follow closely the goals of the experimental design, namely, that in the experiment, rich subjects may employ **FI** to 1) incentivize poor subjects to invest in **ER**, this way 2) encouraging them to contribute in subsequent periods.

In line with SI Note 2.2 above, we capture the first effect by defining a rich subject's strategy as $(C_{R,1}, \dots, C_{R,T}, I_1, \dots, I_T, D_{R,1}, D_{R,T-1})$, where I_t is the contribution to **FI** in period t. Specifically, poor subject j receives $I_t C_{Pj,t}/\sum_{k=1}^{5} C_{Pk,t}$ in period t. Thus, a defective poor subject who contributes 0 does not receive anything. In addition, we assume that no more contributions to **FI** take place when the group reaches the target, because in this case no further contribution is needed.

To capture the second effect, we define a poor subject j's strategy as $(C_{Pj,1}, \dots, C_{Pj,T}, D_{Pj,1}, D_{Pj,T-1}, L_{Pj,1}, \dots, L_{Pj,T})$. Like $D_{Pj,t}$, $L_{Pj,t}$ is also a contribution threshold. Specifically, subject j contributes $C_{Pj,t}$ in period t only if the total **ER** in the previous $t-1$ periods in not less than $D_{Pj,t-1}$ and the total amount contributed in the **FI** stage in period t, I_t, is not less than $L_{Pj,t}$. If $\sum_{k=1}^{t}(C_{R,k}/s + \sum_{j=1}^{5} C_{Pj,k}) < D_{Pj,t}$ or $I_t < L_{Pj,t}$, then the poor subject will contribute 0 in period t and all later periods.

The defective **SPNE** $(\mathbf{C}_P^0, \mathbf{C}_R^0)$ (denoted by $C_{Pj,t} = 0$ and $C_{R,t} = 0$ for all j and t) and the cooperative **QSPNE** (denoted by $\sum_{t=1}^{T} C_{Pj,t}^* = \hat{C}_P^*$, $I_t^* = 0$, $\sum_{t=1}^{T} C_{R,t}^* = \hat{C}_R^*$ and $L_{Pj,t}^* = 0$ for all j and t from Theorem 1) in Control remain **SPNE** in Treatment. Besides these two classes of **SPNE**, Treatment also exhibits a new class of cooperative **QSPNE** that we designate as incentive **QSPNE** $(\mathbf{C}_P^{**}, \mathbf{C}_R^{**})$, where only poor subjects contribute to **ER** and the rich subject incentivizes poor subjects by contributing to **FI**. Specifically, a rich subject's strategy at an incentive **QSPNE** has the form $\mathbf{C}_R^{**} = (0, \dots, 0, I_1^{**}, \dots, I_T^{**}, D_{R,1}^{**}, D_{R,T-1}^{**})$ with $I_t^{**} > 0$ for all $t = 1, \dots, T$. In contrast, a poor subject at an incentive **QSPNE** should not contribute the maximum possible in the early periods of the game, because they will not get any incentive once the group attains the goal. Furthermore, we assume that poor subjects will stop contributing once the **ER** target is reached. Therefore, the total contribution of a poor subject at an incentive **QSPNE** should be $D = 2(1/s + 1)T$, and the **QSPNE** strategy for a poor subject can be denoted by $\mathbf{C}_P^{**} = \left(C_{P,1}^{**}, \dots, C_{P,T}^{**}, D_{P,1}^{**}, D_{P,T-1}^{**}, L_{P,1}^{**}, \dots, L_{P,T}^{**}\right)$ with $\sum_{t=1}^{T} C_{P,t}^{**} = D$ and $C_{P,t}^{**} > 0$ for all $t = 1, \dots, T$.

Let us now discuss for which values of $\mathbf{C}_R^{**}$ and $\mathbf{C}_P^{**}$ the strategy pair of poor and rich subjects is a **QSPNE**.

Theorem S2. A strategy profile $(\mathbf{C}_P^{**}, \mathbf{C}_R^{**})$ with $C_{R,t}^* = 0$, $C_{P,t}^{**}, I_t^{**} > 0$ for all $t = 1, \dots, T$ is a **SPNE** for the Treatment game if (i) $5\sum_{t=1}^{T} C_{P,t}^{**} = D$, (ii) $D_{R,t}^{**}, D_{P,t}^{**} \leq 5\sum_{k=1}^{t} C_{P,k}^{**}$ for all $t = 1, \dots, T-1$, (iii) $I_t^{**} = L_{P,t}^{**}$ for all $t = 1, \dots, T$, (iv) $\sum_{t=1}^{T} I_t^{**} \leq r_R E_R$, $\sum_{t=1}^{T}(C_{P,t}^{**} - I_t^{**}/5) \leq r_P E_P$, and $I_t^{**} \leq 5C_{P,t}^{**}$ for all $t = 1, \dots, T$. At this equilibrium, payoffs for poor and rich subjects are $f_{Pj}(\mathbf{C}_P^{**}, \mathbf{C}_R^{**}) = E_P - \sum_{t=1}^{T}(C_{P,t}^{**} - I_t^{**}/5)$ and $f_R(\mathbf{C}_P^{**}, \mathbf{C}_R^{**}) = E_R - \sum_{t=1}^{T} I_t^{**}$, respectively.

Proof: The proofs for Conditions (i) and (ii) are similar to the proofs for Theorem S1 (i) and (ii). We next prove conditions (iii) and (iv).

(iii) On the one hand, if $I_t^{**} < L_{P,t}^{**}$, i.e., the incentive in period t does not meet the contribution threshold, and thus poor subjects will not contribute in period t and later periods. From condition (i), the group will fail to reach the target. In this case, the rich subject can get a higher payoff by deviating to the defective strategy $\mathbf{C}_R^0$.

On the other hand, if $I_t^{**} > L_{P,t}^{**}$, i.e., the incentive in period t is higher than the contribution threshold, the rich subject can get a higher payoff by decreasing I_t^{**} to $L_{P,t}^{**}$.

Thus, at an incentive **SPNE**, we must have $I_t^{**} = L_{P,t}^{**}$ for all $t = 1, \dots, T$.

(iv) Under conditions (i), (ii), and (iii), the total contributions of a poor subject and a rich subject at $(\mathbf{C}_P^{**}, \mathbf{C}_R^{**})$ are $\sum_{t=1}^{T} C_{P,t}^{**}$ and $\sum_{t=1}^{T} I_t^{**}$, respectively. Thus, the payoffs for a poor subject and a rich subject are $f_{Pj}(\mathbf{C}_P^{**}, \mathbf{C}_R^{**}) = E_P - \sum_{t=1}^{T}(C_{P,t}^{**} - I_t^{**}/5)$ and $f_R(\mathbf{C}_P^{**}, \mathbf{C}_R^{**}) = E_R - \sum_{t=1}^{T} I_t^{**}$, respectively.

We first suppose that the rich subject deviates from the GRIM in period t. Clearly, this player no incentive to increase I_t^{**} because this cannot lead to a higher payoff. Thus, we only need to consider that this player decreases I_t^{**} to 0. In this case, poor subjects will not contribute and the group will fail to meet the target. The maximum expected payoff for the rich subject would be $f_R(\mathbf{C}_P^{**}, \mathbf{C}_R') = (1 - r_R)\left(E_R - \sum_{k=1}^{t-1} I_k^{**}\right)$. Thus, the rich subject will not deviate in period t if $f_R(\mathbf{C}_P^{**}, \mathbf{C}_R^{**}) \geq f_R(\mathbf{C}_P^{**}, \mathbf{C}_R')$, which is equivalent to $r_R E_R \geq r_R \sum_{k=1}^{t-1} I_k^{**} + \sum_{k=t}^{T} I_k^{**}$.

Now suppose that a poor subject j deviates from the GRIM in period t. If this player increases $C_{P,t}^{**}$ to $C_{Pj,t}'$, he/she will get $I_t^{**} C_{Pj,t}'/(C_{Pj,t}' + 4C_{P,t}^{**})$ from **FI**. Thus, the payoff change is

$$\begin{aligned} & f_{Pj}(\mathbf{C}_P^{**}, \mathbf{C}_R^{**}) - f_{Pj}\left(\mathbf{C}_{Pj}' \middle| \mathbf{C}_P^{**}, \mathbf{C}_R^{**}\right) \\ &= -C_{P,t}^{**} + I_t^{**}/5 - \left(-C_{Pj,t}' + I_t^{**} C_{Pj,t}/\left(C_{Pj,t} + 4C_{P,t}^{**}\right)\right) \\ &= (C_{Pj,t}' - C_{P,t}^{**})\left[1 - I_t^{**}/5\left(C_{Pj,t}/4 + C_{P,t}^{**}\right)\right]. \end{aligned} \tag{S5}$$

Notice that $f_{Pj}(\mathbf{C}_P^{**}, \mathbf{C}_R^{**}) - f_{Pj}\left(\mathbf{C}_{Pj}' \middle| \mathbf{C}_P^{**}, \mathbf{C}_R^{**}\right) > 0$ for $I_t^{**} \leq 5C_{P,t}^{**}$; increasing $C_{P,t}^{**}$ cannot lead to a higher payoff. In contrast, if subject j decreases $C_{P,t}^{**}$ to 0, the group will fail to meet the target and her/his maximum expected payoff is $f_{Pj}\left(\mathbf{C}_{Pj}' \middle| \mathbf{C}_P^{**}, \mathbf{C}_R^{**}\right) = (1 - r_P)\left(E_P - \sum_{k=1}^{t-1}(C_{P,k}^{**} - I_k^{**}/5)\right)$. Thus, the poor subject j will not decrease the contribution in period t if $f_{Pj}(\mathbf{C}_P^{**}, \mathbf{C}_R^{**}) \geq f_{Pj}\left(\mathbf{C}_{Pj}' \middle| \mathbf{C}_P^{**}, \mathbf{C}_R^{**}\right)$, which is equivalent to $r_P E_P \geq r_P \sum_{k=1}^{t-1}(C_{P,k}^{**} - I_k^{**}/5) + \sum_{k=t}^{T}(C_{P,k}^{**} - I_k^{**}/5)$.

By the one-shot deviation principle, $(\mathbf{C}_P^{**}, \mathbf{C}_R^{**})$ is a **SPNE** if all players have no incentive to deviate in any period. This requires $r_R E_R \geq r_R \sum_{k=1}^{t-1} I_k^{**} + \sum_{k=t}^{T} I_k^{**}$, $I_t^{**} \leq 5C_{P,t}^{**}$, and $r_P E_P \geq r_P \sum_{k=1}^{t-1}(C_{P,k}^{**} - I_k^{**}/5) + \sum_{k=t}^{T}(C_{P,k}^{**} - I_k^{**}/5)$ for all $t = 1, \dots, T$. Thus, $(\mathbf{C}_P^{**}, \mathbf{C}_R^{**})$ is a **SPNE** if $\sum_{t=1}^{T} I_t^{**} \leq r_R E_R$, $\sum_{t=1}^{T}(C_{P,t}^{**} - I_t^{**}/5) \leq r_P E_P$, and $I_t^{**} \leq 5C_{P,t}^{**}$ for all $t = 1, \dots, T$. □

Since the payoffs of poor and rich subjects at a **SPNE** depend only on $\sum_{t=1}^{T} C_{P,t}^{**}$

and $\sum_{t=1}^{T} I_t^{**}$, we may characterize their strategies by parameters $\hat{C}_{Pj} = \sum_{t=1}^{T} C_{Pj,t}$, $\hat{C}_R = \sum_{t=1}^{T} C_{R,t}$, and $\hat{I} = \sum_{t=1}^{T} I_t$. With these notations, the strategies for poor and rich subjects at the incentive **QSPNE** can be denoted by $\hat{C}_P^{**} = 2(1/s + 1)T$, $\hat{C}_R^{**} = 0$, and $\hat{I}^{**}$ with $10(1 - 2r_P + 1/s)T \leq \hat{I}^{**} \leq 20r_R T$.

Supplemental Note 4. Evolutionary stability of Nash equilibria

4.1. The Control Case

Following SI Note 3 above, we now discuss the evolutionary stability of the equilibria found there. In Control we identified, besides the defective **SPNE**, a class of cooperative **QSPNE.** Specifically, Control 1, 2, 5, and 6 have a unique cooperative **QSPNE** $(\hat{C}_P^*, \hat{C}_R^*) = (2T, 10T)$, while Control 3 and 4 can have multiple equilibria. In this section, we focus on the cooperative **QSPNE** $(\hat{C}_P^*, \hat{C}_R^*) = (2T, 10T)$ and the defective **SPNE** $(\hat{C}_P^0, \hat{C}_R^0) = (0,0)$.

Suppose that poor subjects have two strategies A_1: $\hat{C}_P = 2T$ and A_2:$\hat{C}_P = 0$, and rich subjects have two strategies B_1: $\hat{C}_R = 10T$ and B_2:$\hat{C}_R = 0$. Thus, (A_1, B_1) and (A_2, B_2) correspond to the cooperative **QSPNE** and the defective **SPNE**, respectively. Let us denote the frequencies of strategies A_1 and B_1 in poor population and rich population by x and y, respectively. At each time step, 1 rich subject and 5 poor subjects are randomly chosen to form a 6-person T-period Threshold Public Goods Dilemma. A poor subject using strategy A_2 can always get $4(1 - r_P)$ each period no matter the group composition. But a poor subject using strategy A_1 gets 2 each period if and only if the other 4 poor subjects are also using strategy A_1 and the rich subject is using strategy B_1. To be precise, the single-period expected payoffs for a poor subject using strategy A_1 has the form

$$\hat{f}_P(A_1; x, y) = \sum_{k=0}^{4} \binom{4}{k} x^k (1 - x)^{4-k} [y\hat{f}_P(A_1; k, B_1) + (1 - y)\hat{f}_P(A_1; k, B_2)], \quad (S6)$$

where $\hat{f}_P(A_1; k, B_j)$ is the single-period expected payoff for a poor subject playing A_1 in a group with k other poor subjects playing A_1 and the rich playing B_j. Consequently, $\hat{f}_P(A_1; k, B_j) = 2$ only if $k = 4$ and $j = 1$; otherwise $\hat{f}_P(A_1; k, B_j) = 1$. Thus,

$$\begin{aligned} \hat{f}_P(A_1; x, y) &= \sum_{k=0}^{3} \binom{4}{k} x^k (1 - x)^{4-k} [y + (1 - y)] + 2x^4 y + x^4 (1 - y) \\ &= \sum_{k=0}^{4} \binom{4}{k} x^k (1 - x)^{4-k} + x^4 y = 1 + x^4 y. \end{aligned} \quad (S7)$$

Analogously, a rich subject using strategy B_2 can always get $20(1-r_R)$ each period, and a rich subject using strategy B_1 gets 10 each period only if all the 5 poor subjects are using strategy A_1. The single-period expected payoffs for a rich subject using strategy B_1 has the form

$$\hat{f}_R(B_1;x) = \sum_{k=0}^{5}\binom{5}{k}x^k(1-x)^{5-k}\hat{f}_R(B_1;k), \tag{S8}$$

where $\hat{f}_R(B_1;k)$ is the single-period expected payoff for a rich subject playing B_1 in a group with k poor subjects playing A_1. Specifically, $\hat{f}_R(B_1;k) = 10$ for $k = 5$ and $\hat{f}_R(B_1;k) = 5$ for $k = 0,1,2,3,4$. Thus,

$$\begin{aligned}\hat{f}_R(B_1;x) &= \sum_{k=0}^{4}\binom{5}{k}5x^k(1-x)^{5-k} + 10x^5 \\ &= \sum_{k=0}^{5}\binom{5}{k}5x^k(1-x)^{5-k} + 5x^5 = 5 + 5x^5. \end{aligned}\tag{S9}$$

The replicator dynamics equations for this asymmetric game read

$$\begin{aligned}\frac{dx}{dt} &= x(1-x)(1+x^4y-4(1-r_P)),\\ \frac{dy}{dt} &= 5y(1-y)(1+x^5-4(1-r_R)),\end{aligned}\tag{S10}$$

where both the cooperative **QSPNE** $(x,y) = (1,1)$ and the defective **SPNE** $(x,y) = (0,0)$ are equilibria of the replicator dynamics.

In Control 1, 2, 5, and 6, the payoffs of the cooperative strategies A_1 and B_1 are almost always lower than those associated with the defective strategies A_2 and B_2, i.e., the cooperative **QSPNE** (A_1, B_1) is (weakly) dominated by the defective **SPNE** (A_2, B_2). In these cases, the defective **SPNE** $(x,y) = (0,0)$ is the only stable equilibrium and any initial state with $x \neq 1$ and $y \neq 1$ will converge to it (left panel in Figure S1). In Control 3, strategy B_1 is dominated by B_2. Therefore, the defective **SPNE** $(x,y) = (0,0)$ is the only stable equilibrium (center panel in Figure S1). Finally, in Control 4, the cooperative **QSPNE** $(x,y) = (1,1)$ and the defective **SPNE** $(x,y) = (0,0)$ are both stable, exhibiting a narrow basin of attraction towards to the cooperative **QSPNE** (right panel in Figure S1). These results confirm that increasing the risks of failure for both rich and poor subjects promote **ER**.

4.2. Treatment

As shown in SI Note 3 above, we identified Treatment has three classes of **SPNE** in treatment: The cooperative **QSPNE**, the defective **SPNE** and the incentive **QSPNE**. Since the cooperative **QSPNE** is dominated by the defective **SPNE** in Control 1, 2, 5,

and 6 and is unstable in Control 3, we shall only consider here the competition between the defective **SPNE** and the incentive **QSPNE**. We keep here the same level of analysis carried out in the Control.

Suppose that poor subjects have the following two strategies A_1: $\hat{C}_P = 2(1/s + 1)T$ and A_2: $\hat{C}_P = 0$. While rich subjects have the following two strategies B_1: $(\hat{C}_R, \hat{I}) = (0, \hat{I}^{**})$ and B_2: $(\hat{C}_R, \hat{I}) = (0,0)$. Let us denote the frequencies of strategies A_1 and B_1 in poor population and rich population, respectively, by x and y. A poor subject using strategy A_2 can always get $4(1 - r_P)$, which is independent of x and y. In contrast, a poor subject using strategy A_1 will contribute on average $2(1/s + 1)$ each period if the rich subject is using strategy B_1, and do not contribute if the rich subject is using strategy B_2. In contrast, a rich subject using strategy B_1 will contribute $I^{**} = \hat{I}^{**}/T$ each period. Thus, the single-period expected payoffs for a poor subject using strategy A_1 is given by

$$\hat{f}_P(A_1; x, y) = \sum_{k=0}^{4} \binom{4}{k} x^k (1-x)^{4-k} [y\hat{f}_P(A_1; k, B_1) + (1-y)\hat{f}_P(A_1; k, B_2)], \quad \text{(S11)}$$

where $\hat{f}_P(A_1; k, B_1) = 2(1 - 1/s) + I^{**}/5$ for $k = 4$, $\hat{f}_P(A_1; k, B_1) = 1 - 1/s + I^{**}/(2 + 2k)$ for $k = 0,1,2,3$ (i.e., $k + 1$ poor subjects using strategy A_1 share the contribution I^*), and $\hat{f}_P(A_1; k, B_2) = 4(1 - r_P)$ for all k. We obtain

$$\begin{aligned}\hat{f}_P(A_1; x, y) &= \sum_{k=0}^{3} \binom{4}{k} x^k (1-x)^{4-k} [y(1 - 1/s + I^{**}/(2 + 2k)) + 4(1 \\ &\quad - r_P)(1-y)] + x^4 y(2(1 - 1/s) + I^{**}/5) + 4(1 - r_P)x^4(1-y) \\ &= 4(1 - r_P)(1-y) + (1 - 1/s)y(1 + x^4) + I^{**}y[1 + x^5 - (1-x)^5]/(10x).\end{aligned} \quad \text{(S12)}$$

Analogously, a rich subject using strategy B_2 can always get $20(1 - r_R)$ each period, and the single-period expected payoff for a rich subject using strategy B_1 depends on the number of poor subjects using strategy A_2, which is given by

$$\hat{f}_R(B_1; x) = \sum_{k=0}^{5} \binom{5}{k} x^k (1-x)^{5-k} \hat{f}_R(B_1; k), \quad \text{(S13)}$$

where $\hat{f}_R(B_1; k) = 20 - I^{**}$ for $k = 5$, and $\hat{f}_R(B_1; k) = 10 - I^{**}/2$ for $k = 0,1,2,3,4$. Thus, we obtain

$$\hat{f}_R(B_1; x) = 10 - I^{**}/2 + [10 - I^{**}/2]x^5. \quad \text{(S14)}$$

The resulting replicator dynamics equations read

$$\frac{dx}{dt} = x(1-x)y[-4(1 - r_P) + (1 - 1/s)(1 + x^4) + I^{**}[1 + x^5 - (1-x)^5]/(10x)],$$

$$\frac{dy}{dt} = y(1-y)[10(2r_R - 1) - I^{**}/2 + [10 - I^{**}/2]x^5],$$

(S15)

The incentive **QSPNE** $(x, y) = (1,1)$ is locally asymptotically stable if $10(1 - 2r_P + 1/s) \leq I^{**} \leq 20r_R$, and the defective **SPNE** $(x, y) = (0,0)$ is unstable if $6 - 8r_P + 2/s < I^{**}$. Thus, larger r_R and r_P lead to an increase in the stability of the incentive **QSPNE**. It is worth noting that, Eq.(S15) is independent of T, which means that, at this level of analysis, the effectiveness of **FI** is not affected by the number of periods but depends on the average incentive per period I^{**}. From the first equation of Eq.(S15), the direction of evolution of x is independent of y. Furthermore, increasing the I^{**} and r_P will promote an increase of contributions by poor subjects. For small I^{**} and r_P, $\frac{dx}{dt} < 0$ for all x. For intermediate I^{**} and r_P, there may exist $0 < x_1^* < x_2^* < 1$ such that $\frac{dx}{dt} > 0$ for $x < x_1^*$ and $x > x_2^*$, and $\frac{dx}{dt} < 0$ for $x_1^* < x < x_2^*$. For large I^{**} and r_P, $\frac{dx}{dt} > 0$ for all x. On the other hand, the second equation indicates that $\frac{dy}{dt} > 0$ for $x > x^{**} = \left(\frac{I^{**} - 20(2r_R - 1)}{20 - I^{**}}\right)^{1/5}$ with x^{**} an increasing function of I^{**} and a decreasing function of r_R. Thus, as shown in SI Figure S2, most trajectories starting from $x > x^{**}$ will converge to the incentive **QSPNE** for $10(1 - 2r_P + 1/s) \leq I^{**} \leq 20r_R$.

Supplemental Note 5. Data analysis

SI Table S1 shows the total **ER** and average relative contributions of rich and poor subjects in different groups. In Control and Treatment with (r_R, r_P)=(0.5,0.5), the differences obtained for the total **ER** when comparing experiments with s=2 and s=3 as well as T=6 and T=10 are not statistically significant. The same occurs when we compare the contributions of the poor to **ER** and of the rich to **FI** (see SI Table S2). Similarly, in Control and Treatment sessions with (T,s)=(6,3), the differences in total **ER** among experiments with (r_R, r_P) = (0.5,0.5), (0.5,0.7), and (0.7,0.7), respectively, are not statistically significant. Furthermore, only increasing the risk of poor subjects does not affect the behaviours of the rich and poor, but increasing the risk of rich subjects could encourage them to contribute more to **ER** in Control and **FI** in Treatment (see SI Table S3).

Altogether, the total **ER** in Control is significantly below the target, whereas the total **ER** in Treatment is not statistically different from the target (see SI Table S1). These results show that **FI** constitutes a robust mechanism of promoting **ER**.

Specifically, poor subjects contribute significantly more in Treatment compared to Control (two-side Mann-Whitney U-test, p-values <0.001). In contrast, rich subjects contribute significantly less to **ER** in Treatment compared to Control (two-side Mann-Whitney U-test, p-values <0.001), but the total contributions (i.e., **ER+FI**) are significantly higher compared to Control (two-side Mann-Whitney U-test, p-values <0.001). Finally, the wealth of poor subjects at the end of the game in Treatment is significantly higher compared to Control, which means that **FI** effectively reduces the net **ER** cost of poor subjects (two-side Mann-Whitney U-test, p-values <0.001).

Supplemental Note 6. Robustness test

To test the robustness of the results, namely, to which extent the presence of more than 1 rich subject in each group may affect the performance of **FI**, we carried out 2 additional sessions of experiments (Control 2R and Treatment 2R), involving 5 groups of 12 subjects. In the Control 2R and Treatment 2R sessions, each group has 2 rich subjects and 10 poor subjects. Results of the experiments show that having more than one rich subject in the group will contribute to "dilute" each one's responsibility, rendering coordination towards the goal more difficult. As shown in SI Table S4, all the 5 Control groups failed, and only 1 out of the 5 Treatment groups successfully reached the target. Although adding the **FI** still improves both **ER** and the rates of success, both quantities are lower compared to groups of half the size and with 1 rich subject only. Besides the group size effect that also contributes to render more difficult the probability of success, it is worth pointing out that, in the only group that reached the target under Treatment, both rich subjects contributed about half of their endowment to **FI**. On the other hand, in failed groups, rich subjects either free-rode or contributed more to **ER** rather than **FI** (SI Figure S5). This suggests that it is crucial that rich subjects assume their role and responsibility for the success of global **ER**.

Supplemental Figures

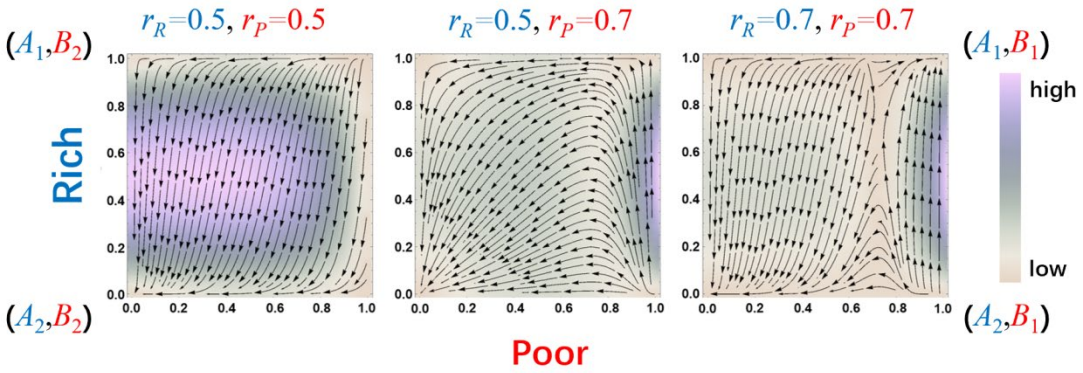


Figure S1. Control: Evolutionary dynamics of (Q)SPNE Strategies. The arrows indicate the direction of the gradient of selection whose projections, at any point in the phase space (x,y) — where x represents the fraction of poor subjects playing strategy B_1 =2T and y the fraction of the rich playing strategy A_1 =10T — are given by the right hand sides of SI Eq.(S10). Each panel corresponds to different values for r_R and r_P as indicated. The colour scale shown applies to the rate of change of the evolutionary dynamics that fills the background of the plot. Note that, in this analysis, group compositions always have a single rich subject and 5 poor subjects. The defective **SPNE** (A_2, B_2) is globally stable in the left and the center panels. However, as risk goes up (both for the poor and for the rich) the stability of this defective **SPNE** weakens, and the cooperative **QSPNE** becomes stable in the right panel.

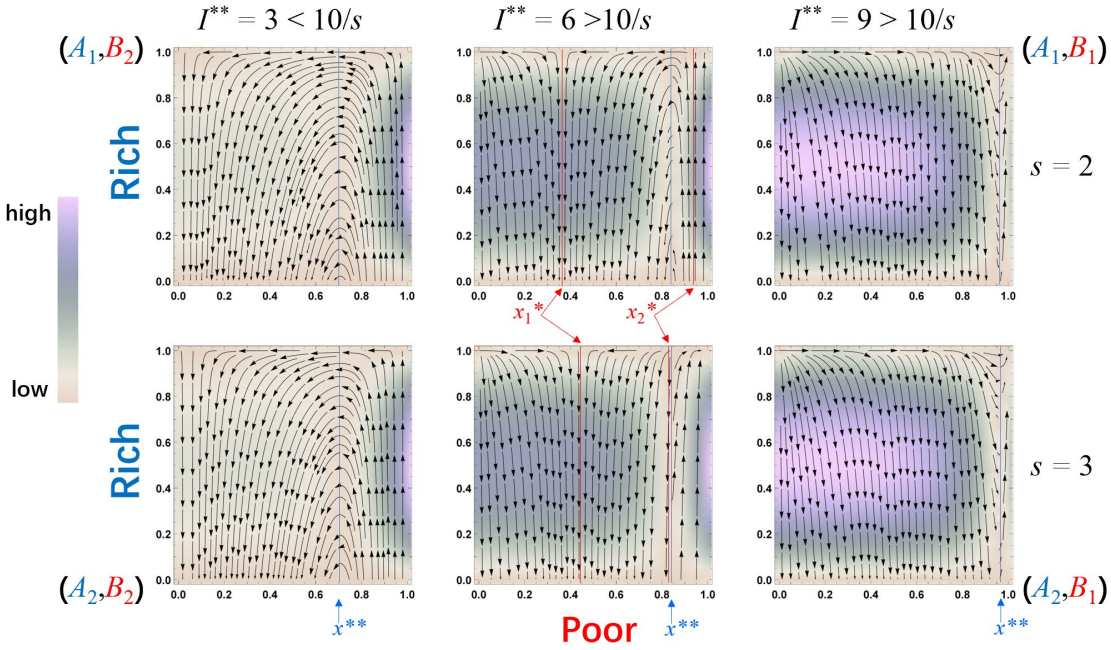


Figure S2. Treatment: Evolutionary dynamics of (Q)SPNE Strategies. We use the same conventions of SI Figure S1 (in all cases $r_R = r_P = 0.5$). Each corner of the phase space corresponds to a pair of strategies. The defective **SPNE** $(A_2, B_2) = (0,0)$ is always locally stable. For the average incentive per period $I^{**} > 10/s$, the incentive **QSPNE** (See Methods and SI Eq.(S15)) becomes locally stable.

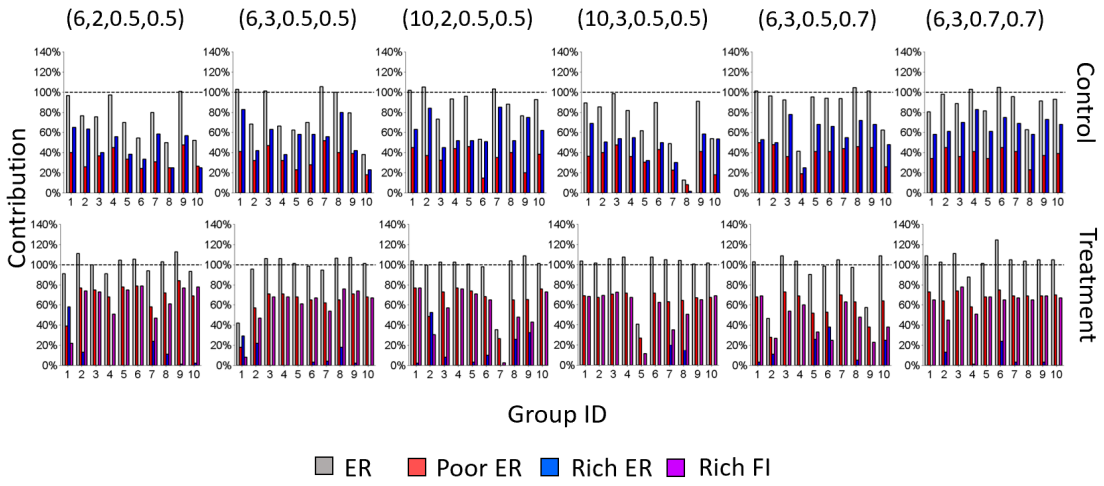


Figure S3. Percentage values for total ER and contributions to ER and FI of each group. We use the same notation and conventions as in Figure 2 and Figure 3 of main text. Upper panels: Control. Lower panels: Treatment.

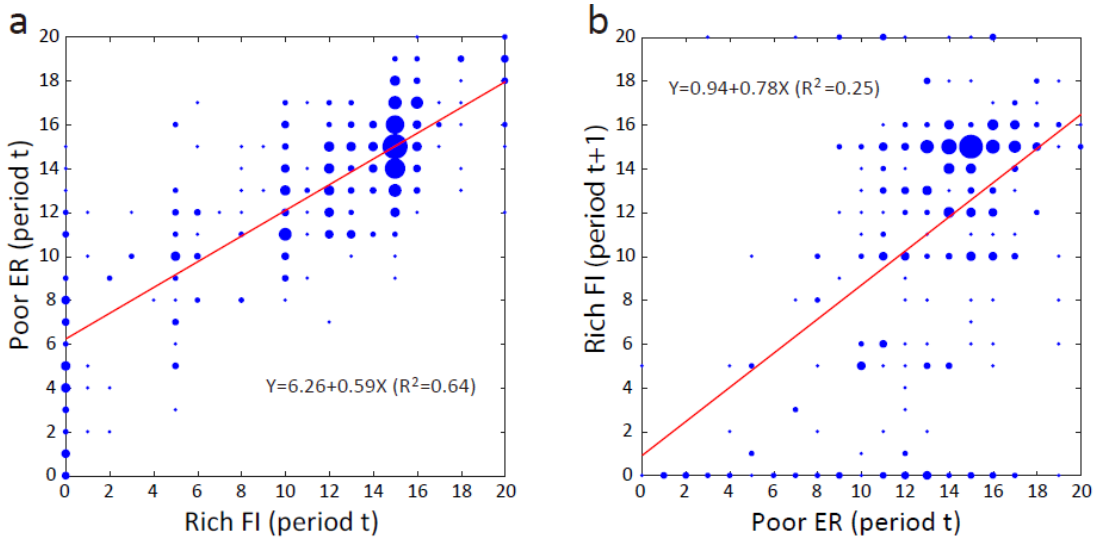


Figure S4. Correlation between ER of the poor and FI of the rich at the group level. We analyze the correlation between the contribution to **ER** by the poor and the contribution to **FI** by the rich through linear regression and Pearson correlation. **(a)** The correlation between the contribution of rich subjects to **FI** in the **FI** stage in period t and the total contributions of poor subjects to **ER** in the **ER** stage of the same period. ($Y=6.26+0.59X$, F-test, $R^2=0.64$, p-value<0.001; Pearson correlation coefficient is 0.7986). (b) The correlation between the total contributions of poor subjects to **ER** in the **ER** stage in period t and the contribution of rich subjects to **FI** in the **FI** stage in period $t+1$ ($Y=0.94+0.78X$, F-test, $R^2=0.25$, p-value<0.001; Pearson correlation coefficient is 0.5015). Comparisons between regression results and Pearson correlation coefficients for the two situations suggest that it is more reasonable that the **FI** starts being smaller first and then the poor **ER** follows.

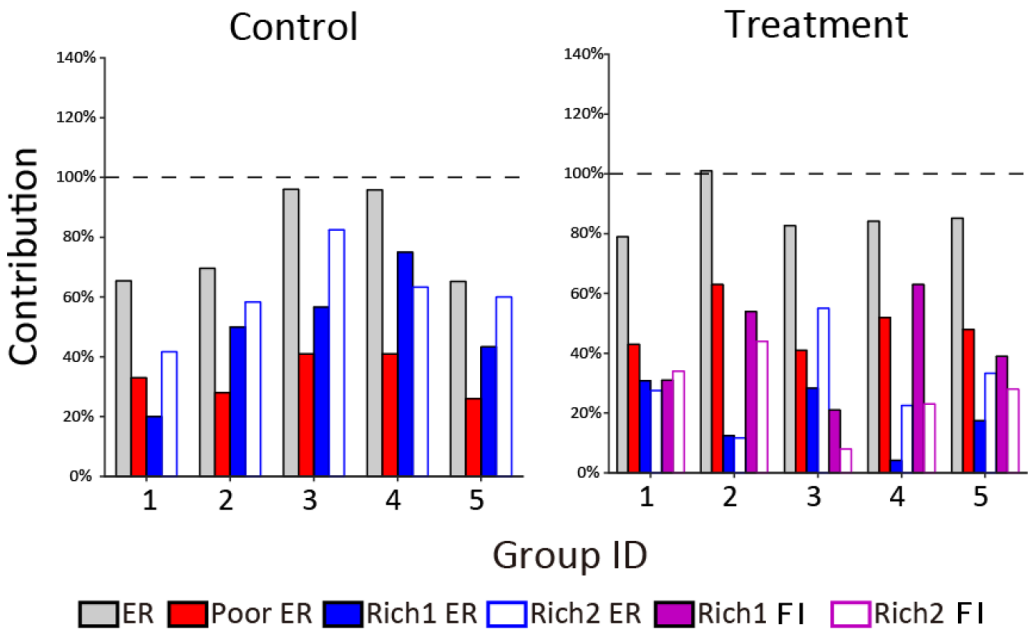


Figure S5. Average percentage values for total ER and contributions to ER and FI in Control 2R and Treatment 2R. We use the same notation and conventions as in Figure 2 and Figure 3 of main text. Left panel: Control. Right panel: Treatment.

Figure S6. Screen shots of experimental interfaces. All screenshots refer to Treatment with T=10-period, used here as an example. Interface of rich subjects in the **FI** stage (top) and in the **ER** stage (middle), and poor subjects in the **ER** stage (bottom). Left panels provide screenshots of the original interface (in Chinese), whereas right panels provide direct English translations.

Supplemental Tables

Type Control/Treatment (T, s, r_R, r_P)	Total **ER**	Poor			Rich		
		Contribution in **ER**	**FI** received	Wealth at end	Contribution in **ER**	Contribution in **FI**	Wealth at end
Control 1 (6, 2, 0.5, 0.5)	75%**	34%	-	67%	46%	-	54%
Treatment 1 (6, 2, 0.5, 0.5)	100%	70%	64%	94%	11%	64%	25%
Control 2 (6, 3, 0.5, 0.5)	79%*	35%	-	65%	54%	-	46%
Treatment 2 (6, 3, 0.5, 0.5)	96%	61%	59%	97%	8%	59%	33%
Control 3 (6, 3, 0.5, 0.7)	88%*	40%	-	60%	58%	-	42%
Treatment 3 (6, 3, 0.5, 0.7)	92%	58%	44%	86%	11%	44%	45%
Control 4 (6, 3, 0.7, 0.7)	90%*	37%	-	63%	68%	-	32%
Treatment 4 (6, 3, 0.7, 0.7)	105%	69%	64%	95%	4%	64%	32%
Control 5 (10, 2, 0.5, 0.5)	88%*	35%	-	65%	62%	-	39%
Treatment 5 (10, 2, 0.5, 0.5)	96%	65%	54%	89%	13%	54%	33%
Control 6 (10, 3, 0.5, 0.5)	71%**	32%	-	68%	45%	-	55%
Treatment 6 (10, 3, 0.5, 0.5)	98%	64%	57%	93%	3%	57%	40%
Control 1 to 6	82%**	36%		65%	56%		44%
Treatment 1 to 6	98%	65%	57%	92%	8%	57%	35%

Table S1. Total ER and average relative contributions in different types of experiments. The second column gives the percentage values (relative to the target **ER**) in each game. The statistical results of two-side Wilcoxon signed rank test indicate that the total **ER** in Treatment groups is not significantly different from the target. P-value<0.05 and <0.01 are indicated by * and **, respectively. Values provided for the poor and for the rich are represented as the fraction of their initial endowments.

$(r_R, r_P) = (0.5,0.5)$	Type	(6,2) vs (6,3)	(10,2) vs (10,3)	(6,2) vs (10,2)	(6,3) vs (10,3)
Total ER	Control	0.623	0.089	0.151	0.427
	Treatment	0.791	0.384	0.910	0.520
Poor ER	Control	0.705	0.650	0.521	0.910
	Treatment	0.045	0.345	0.385	0.570
Rich ER	Control	0.449	0.121	0.121	0.344
	Treatment	1.000	0.052	0.616	0.114
Rich FI	Treatment	0.325	0.940	0.241	1.000

Table S2. p-values of two-side Mann-Whitney U-test. In each of Control and Treatment with $(r_R, r_P)=(0.5,0.5)$, the differences in total **ER** in experiments with s=2 and 3 as well as T=6 and 10 are not statistically significant, the same occurring for the contributions of the poor and the rich to **ER** and of the rich to **FI**.

(T, s) = (6, 3)	Type	(0.5,0.5) vs (0.5,0.7)	(0.5,0.5) vs (0.7,0.7)	(0.5,0.7) vs (0.7,0.7)
Total ER	Control	0.650	0.406	0.734
	Treatment	0.880	0.880	0.150
Poor ER	Control	0.272	0.569	0.254
	Treatment	0.471	0.160	0.048*
Rich ER	Control	0.623	0.041*	0.112
	Treatment	0.668	0.668	0.284
Rich FI	Treatment	0.059	1.000	0.010*

Table S3. p-values of two-side Mann-Whitney U-test. In each of Control and Treatment groups with (T, s)=(6, 3), the differences in total **ER** in experiments among (r_R, r_P) = (0.5,0.5), (0.5,0.7), and (0.7,0.7) are not statistically significant. In addition, when both rich and poor subjects have higher risks, rich subjects contribute more to **ER** in Control, and rich subjects contribute more to **FI** and poor subjects contribute more to **ER** in Treatment. P-value<0.05 is indicated by *.

Type (T, s, r_R, r_P)	Successful groups/ Total groups	Total **ER**	Poor			Rich		
			Contribution in **ER**	**FI** received	Wealth at end	Contribution in **ER**	Contribution in **FI**	Wealth at end
Control 2R (6,3,0.5,0.5)	0/5	78%	34%	-	66%	55%	-	45%
Treatment 2R (6,3,0.5,0.5)	1/5	86%	50%	33%	83%	24%	33%	43%

Table S4. Total ER and average relative contributions in Control 2R and Treatment 2R. The second column shows the number of successful groups and the total number of groups employed. The third column gives the percentages relative to the target **ER** values in each game. Values provided for the poor and for the rich are represented as the fraction of their respective initial endowments.

Type and number of groups		Parameters (N_R=1, N_P=5)							(Q)NE
		T	s	E_R	E_P	D	r_R	r_P	
Control 1	10	6	2	120	24	90	0.5	0.5	$(C_R = C_P = 0)$; $(C_R = 10T, C_P = 2T)$
Treatment 1	10								
Control 2	10	6	3	120	24	80	0.5	0.5	$(C_R = C_P = 0)$; $(C_R = 10T, C_P = 2T)$
Treatment 2	10								
Control 3	10	6	3	120	24	80	0.5	0.7	$(C_R = C_P = 0)$; $(C_R = 10T, C_P = 2T)$
Treatment 3	10								
Control 4	10	6	3	120	24	80	0.7	0.7	$(C_R = C_P = 0)$; $(C_R = 10T, C_P = 2T)$
Treatment 4	10								
Control 5	10	10	2	200	40	150	0.5	0.5	$(C_R = C_P = 0)$; $(C_R = 10T, C_P = 2T)$
Treatment 5	10								
Control 6	10	10	3	200	40	133	0.5	0.5	$(C_R = C_P = 0)$; $(C_R = 10T, C_P = 2T)$
Treatment 6	10								

Table S5. Details of experimental settings. Each group consists of 6 subjects, 1 rich subject and 5 poor subjects. E_R and E_P are the initial endowments of a rich and a poor subject, respectively. D is the **ER** target, T is the number of periods, and s is the abatement cost factor for the rich. If a group fails to reach the target, rich and poor subjects will lose all of their savings with probability r_R and r_P, respectively. Let C_R and C_P denote the total contributions of a rich and a poor subject in a T-period game, respectively. All settings are predicted to exhibit similar equilibria structure: A non-cooperative equilibrium, where no one invests (i.e., $C_R = C_P = 0$), and a cooperative equilibrium, where both rich and poor subjects invest half of their endowments (i.e., $C_R = 10T, C_P = 2T$). In addition, Control 3 and 4 can have multiple cooperative equilibria, and Treatment groups also exhibit a class of incentive equilibria, where only the poor invest in **ER** whereas the rich contribute solely to **FI**.

Experiment	Parameters	Average values and standard deviations	Period 1	Period 2	Period 3	Period 4	Period 5	Period 6
Control	(6, 2, 0.5, 0.5)	AV of poor **ER**	0.385	0.390	0.340	0.360	0.315	0.220
		SD of poor **ER**	0.074	0.114	0.116	0.086	0.160	0.155
		AV of rich **ER**	0.525	0.465	0.525	0.440	0.370	0.440
		SD of rich **ER**	0.075	0.187	0.087	0.246	0.313	0.383
	(6, 3, 0.5, 0.5)	AV of poor **ER**	0.395	0.370	0.435	0.295	0.305	0.295
		SD of poor **ER**	0.082	0.121	0.090	0.119	0.154	0.180
		AV of rich **ER**	0.540	0.630	0.555	0.600	0.635	0.285
		SD of rich **ER**	0.092	0.216	0.324	0.265	0.246	0.385
	(6, 3, 0.5, 0.7)	AV of poor **ER**	0.410	0.440	0.395	0.370	0.355	0.400
		SD of poor **ER**	0.111	0.077	0.079	0.133	0.163	0.216
		AV of rich **ER**	0.525	0.550	0.545	0.500	0.655	0.710
		SD of rich **ER**	0.108	0.195	0.207	0.291	0.289	0.311
	(6, 3, 0.7, 0.7)	AV of poor **ER**	0.370	0.385	0.380	0.365	0.345	0.400
		SD of poor **ER**	0.071	0.084	0.090	0.148	0.162	0.116
		AV of rich **ER**	0.520	0.640	0.665	0.770	0.740	0.725
		SD of rich **ER**	0.040	0.118	0.157	0.178	0.296	0.384
Treatment	(6, 2, 0.5, 0.5)	AV of poor **ER**	0.705	0.755	0.675	0.730	0.660	0.675
		SD of poor **ER**	0.140	0.106	0.142	0.219	0.200	0.214
		AV of rich **ER**	0.125	0.105	0.055	0.130	0.120	0.120
		SD of rich **ER**	0.223	0.165	0.091	0.282	0.194	0.239
		AV of rich **FI**	0.640	0.675	0.670	0.680	0.570	0.580
		SD of rich **FI**	0.201	0.178	0.154	0.286	0.286	0.311
	(6, 3, 0.5, 0.5)	AV of poor **ER**	0.630	0.635	0.605	0.675	0.660	0.480
		SD of poor **ER**	0.125	0.192	0.215	0.187	0.203	0.299
		AV of rich **ER**	0.015	0.010	0.080	0.085	0.130	0.145
		SD of rich **ER**	0.045	0.030	0.158	0.148	0.234	0.299
		AV of rich **FI**	0.590	0.635	0.580	0.650	0.635	0.445
		SD of rich **FI**	0.214	0.236	0.212	0.184	0.158	0.378
	(6, 3, 0.5, 0.7)	AV of poor **ER**	0.585	0.555	0.580	0.675	0.580	0.490
		SD of poor **ER**	0.170	0.215	0.183	0.159	0.200	0.228
		AV of rich **ER**	0.105	0.100	0.065	0.135	0.080	0.160
		SD of rich **ER**	0.140	0.116	0.107	0.227	0.114	0.244
		AV of rich **FI**	0.520	0.465	0.450	0.550	0.465	0.175
		SD of rich **FI**	0.198	0.286	0.266	0.285	0.192	0.225
	(6, 3, 0.7, 0.7)	AV of poor **ER**	0.370	0.385	0.380	0.365	0.345	0.400
		SD of poor **ER**	0.071	0.084	0.090	0.148	0.162	0.116
		AV of rich **ER**	0.520	0.640	0.665	0.770	0.740	0.725
		SD of rich **ER**	0.040	0.118	0.157	0.178	0.296	0.384
		AV of rich **FI**	0.655	0.690	0.715	0.770	0.715	0.590
		SD of rich **FI**	0.113	0.170	0.095	0.095	0.074	0.226

Table S6. Details of the 6-period experiments. Average relative contributions and the corresponding standard deviations.

Experiment	Parameters	Average values and standard deviations	Period 1	Period 2	Period 3	Period 4	Period 5	Period 6	Period 7	Period 8	Period 9	Period 10
Control	(10, 2, 0.5, 0.5)	AV of poor **ER**	0.355	0.400	0.405	0.335	0.370	0.345	0.350	0.315	0.335	0.315
		SD of poor **ER**	0.042	0.134	0.115	0.112	0.162	0.160	0.100	0.134	0.182	0.182
		AV of rich **ER**	0.560	0.545	0.530	0.550	0.630	0.515	0.615	0.640	0.895	0.730
		SD of rich **ER**	0.151	0.079	0.100	0.092	0.200	0.276	0.316	0.301	0.152	0.385
	(10, 3, 0.5, 0.5)	AV of poor **ER**	0.410	0.400	0.345	0.325	0.320	0.375	0.325	0.270	0.255	0.205
		SD of poor **ER**	0.099	0.095	0.146	0.135	0.158	0.165	0.152	0.162	0.162	0.171
		AV of rich **ER**	0.490	0.505	0.520	0.520	0.490	0.450	0.355	0.440	0.330	0.440
		SD of rich **ER**	0.308	0.272	0.282	0.346	0.317	0.277	0.320	0.254	0.372	0.444
Treatment	(10, 2, 0.5, 0.5)	AV of poor **ER**	0.680	0.655	0.645	0.605	0.650	0.675	0.660	0.670	0.645	0.625
		SD of poor **ER**	0.105	0.181	0.144	0.217	0.202	0.218	0.200	0.215	0.243	0.276
		AV of rich **ER**	0.075	0.070	0.120	0.100	0.130	0.120	0.100	0.090	0.105	0.430
		SD of rich **ER**	0.098	0.108	0.194	0.184	0.205	0.199	0.167	0.164	0.203	0.415
		AV of rich **FI**	0.535	0.555	0.570	0.530	0.550	0.620	0.565	0.585	0.590	0.330
		SD of rich **FI**	0.237	0.295	0.269	0.261	0.210	0.230	0.249	0.228	0.301	0.354
	(10, 3, 0.5, 0.5)	AV of poor **ER**	0.680	0.655	0.645	0.605	0.650	0.675	0.660	0.670	0.645	0.625
		SD of poor **ER**	0.105	0.181	0.144	0.217	0.202	0.218	0.200	0.215	0.243	0.276
		AV of rich **ER**	0.060	0.050	0.015	0.015	0.030	0.075	0.030	0.040	0.015	0.015
		SD of rich **ER**	0.137	0.100	0.045	0.045	0.090	0.160	0.075	0.083	0.045	0.045
		AV of rich **FI**	0.645	0.710	0.630	0.660	0.645	0.625	0.555	0.570	0.545	0.130
		SD of rich **FI**	0.163	0.077	0.225	0.261	0.282	0.227	0.283	0.292	0.308	0.254

Table S7. Details of the 10-period experiments. Average relative contributions and the corresponding standard deviations.