Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Sanctions and international interaction improve cooperation to avert climate change

Gianluca Grimalda (✉ gianluca.grimalda@ifw-kiel.de)
Kiel Institute for the World Economy https://orcid.org/0000-0002-5605-5591
Alexis Belianin
Higher School of Economics
Heike Hennig-Schmidt
University of Bonn
Till Requate
University of Kiel
Marina Ryzhkova
Tomsk State University

Article

Keywords: climate change, sanctions, international politics

Posted Date: August 16th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-777082/v2

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Version of Record: A version of this preprint was published at Proceedings of the Royal Society B: Biological Sciences on April 6th, 2022. See the published version at https://doi.org/10.1098/rspb.2021.2174.

Abstract

Imposing sanctions on noncompliant parties to international agreements is often advocated as a remedy for international cooperation failure, notably in climate agreements. We provide an experimental test of this conjecture in a collective-risk social dilemma simulating the effort to avoid catastrophic climate change. We involve groups of participants from two cultural areas that were shown to achieve different levels of cooperation nationally when peer-level sanctions were available. Here we show that, while this result still holds nationally, international interaction backed by sanctions is overall beneficial. Cooperation by low cooperator groups increases in comparison with national cooperation and converges to the cooperation levels of high cooperation groups. While such an increase is small without sanctions, it becomes sizable when sanctions are imposed. Revealing or hiding counterparts' nationality does not affect results. Our study supports the proposal to use sanctions to support international cooperation to avert collective risk such as climate change.

Introduction

A wide range of problems facing humanity – such as overexploitation of natural resources and climate change – are global in scale and require international cooperation across widely different cultures[1]. Currently, international cooperation falls critically short of the levels necessary to mitigate climate change[1-3]. While cooperation may be sustained by direct and indirect reciprocity[4] in small or culturally cohesive groups[5], cooperation in large groups of unrelated individuals is typically parochial, that is, it favors others perceived as belonging to one's own group at the expense of others perceived as belonging to other groups[6-8]. Since nationality is one of the strongest sources of parochial attachment[9], international cooperation appears problematic[10]. Some scholars have proposed the introduction of substantial and credible trade sanctions for countries that do not comply with climate agreements as a possible solution to the current stalemate[1,11]. Sanctions could take the form of increased tariffs on imported goods from countries not complying with climate agreements. This type of sanctions are also at the basis of the "climate club" proposal, where countries not complying with a climate agreement suffer a penalty in the form of increased tariffs from countries belonging to the club[1]. Yet, applying sanctions may trigger a second-order cooperation problem[12]. Sanctioning is generally costly to the party applying sanctions, thus each party will prefer to free ride on others' sanctions. Nevertheless, individuals seem inclined to apply sanctions on others even when this is costly to them[13-14], and this also holds for countries[2]. As a matter of fact, the number of climate provisions introduced in trade agreements is increasing[15].

We designed an experiment to test the effectiveness of sanctions for increasing international cooperation in interaction mimicking the costs and incentives that individuals face to prevent climate change. Our experiment builds on the Collective Risk Social Dilemma[16] (CRSD; See Supplementary Information (SI): Supplementary Notes: Section S7 for an abbreviation list). We modify the CRSD by introducing a sanctioning stage that reflects the characteristics of trade sanctions applied to climate agreements.

Controlled experimental evidence on sanctioning in cross-cultural contexts is rare[17] and lacking in international contexts. We involve participants from two countries - Germany and Russia – epitomizing cultural areas where sanctions have been found to work or fail, respectively, as mechanisms to increase cooperation[18-19]. This puts the potential impact of sanctions as a method for underpinning international cooperation to a severe test. Theoretically, it is unclear whether cooperation will increase or decrease when individuals from different cultural backgrounds come together. Experiments indicate the existence of a “bad apple” effect, i.e. the presence of a few low cooperators in a group leads to a drastic reduction of willingness to cooperate in others[20]. This would imply high cooperators lowering their cooperation rates and matching those of low cooperators. On the other hand, if low cooperation is caused by pessimism about others’ willingness to cooperate, this may prompt low cooperators to switch from a strategy based on mistrust and low cooperation to one based on trust in others’ high cooperation[21], when they are matched with high cooperators in an international context.

Scientists classify climate change as having both a “gradual” and a “catastrophic” component. The former refers to incremental changes in underlying factors that continuously alter the climate, such as the progressive rise in sea levels. Catastrophic climate change refers to structural changes in eco-systems triggered by temperatures exceeding a “tipping point” and leading to irreversible change[22]. Examples are the collapse of the Amazon forest or the loss of ice sheets. The CRSD used in our experiments captures in a stylized way the potential gains and losses underlying catastrophic climate change[16]. Groups of individuals are faced with the possibility of losing part of their endowment if a random loss event occurs. To prevent such collective losses, individuals can contribute part of their monetary endowments to a collective fund that reduces the probability of the loss event occurring. Possible losses are large, thus simulating a major catastrophe in the offing. The consensus among scientists is that if temperatures increase less than 2˚C from pre-industrial levels, no catastrophic loss will occur. We call this level the “certain safety threshold”. On the other hand, an increase by more than 5˚C by 2100– which would occur in a “business-as-usual” scenario - will certainly trigger catastrophic climate change[22]. We call this the “certain unsafety threshold”. There is, however, uncertainty over which temperature level will actually trigger catastrophic climate change within the 2˚C–5˚C range[22-23]. We model uncertainty about the actual temperature level associated with this “catastrophe tipping point” using a uniform distribution over the interval bounded by the “certain safety” and the “certain unsafety” thresholds[23]. Collective loss is thus avoided with a probability proportional to the total amount of money that the group invests in the collective fund, relative to the amount of investment needed to achieve the certain safety threshold.

Experiment Design

Participants were involved in the CRSD at either the national (NAT) or the international (INT) level, with sanctions being possible (S-treatments) or not possible (NS-treatments). Groups of six participants interacted in the CRSD, three of whom were university students in one city and three in another. In NAT treatments, the two cities were either in Germany or in Russia. In INT treatments, one city was in Germany and one in Russia. Students were citizens of their country of residence. (See SI: section S1.1, S1.2 for

participants' demographic and cultural characteristics). The international treatments were conducted under two different settings: In the Blind (B) treatments, participants were not made aware that students from the other city were actually from another country[24]. In the Open (O-)treatments, conversely, German and Russian participants were informed that the other city was located either in Russia or in Germany. Students were citizens of their country of residence. Behavior in international interaction may be affected by prejudice and stereotypes about foreigners[10], by national pride, or by the desire to outperform the other group[25]. Our experimental design permits the comparison of outcomes between the case where such prejudices or inter-group motivations may be operative – i.e., in the O-treatments – and the case where prejudice or inter-group motivations are ruled out by construction – i.e., in the B-treatments. Ex-post questionnaire data confirm that our treatment manipulation worked because a large majority of participants in the B-treatments believed that they were interacting with participants from the same country (Table S4). The outcomes of the B-treatments can thus be attributed solely to the effect of participants' choices, ruling out all beliefs and motivations relative to interaction with foreigners. The eight experimental treatments are summarized in Table 1. The treatments were exogenous to demographic characteristics, participants' university degree, proxies of socio-economic status, within each country (Table S5).

Participants interacted over ten periods with the same partners in real-time via the Internet. Interactions were anonymous, but each group member could be identified by a label. Each participant was endowed with 60 tokens in each period. Each token was worth €0.07 in Germany and 2.0 Ruble in Russia. Such levels ensured equivalent purchasing power across countries. In the NS-treatments, participants could contribute up to 50 tokens to a collective fund, the remaining 10 tokens being automatically added to their private accounts. If the sum of total contributions (C) to the collective fund exceeded the certain safety threshold (T), there would be no loss to any player's private account. If, however, C<T at the end of the ten periods, a loss of 75% to each player's private account would occur with probability 1-P, where P=min{C/T;1} and P is the probability of loss avoidance (PLA) (SI: Fig. S4). P was the same for each group member. C=0 is the certain unsafety threshold. Individuals' private accounts at the end of ten periods would equal the total endowment of 600 tokens minus total individual contributions to the collective fund. Participants earned either the full amount in the private accounts at the end of the ten rounds if no loss occurred, or else a quarter of this amount.

The CRSD in the S-treatments took place in two stages. The first stage was identical to the NS-treatments. In the second stage, each group member could use the remaining 10 tokens from their endowments to reduce other group members' private accounts in each of the 10 periods. Tokens spent on such sanctions were deducted from the private account. This sanctioning system had a number of characteristics in common with typical sanctions in international trade agreements. First, sanctions were observable26 as tariff systems are known to all relevant parties. Second, the number of tokens deducted from a punished participant's account increased more than proportionally in the number of tokens spent by other participants to sanction that participant. Similarly, the costs incurred by a country rise disproportionately as the number of sanctioning countries increases and as sanctions become more

severe. The sanctioning cost structure is reported in Table S6. Final payoffs under the S-treatments were equal to those under the NS-treatments minus the sanctioning costs.

We use two theoretical benchmarks to analyze this interaction. The Nash Equilibrium (NE) identifies the set of individual actions ensuring that each action is the best response to others' individual actions assuming that each agent maximizes their own monetary payoff. By contrast, the cooperative solution (CS) takes the perspective of the entire group and maximizes the total sum of expected monetary payoffs (SI: Section S1.3 for the derivation of the two solutions).

For low levels of T, both the NE and the CS prescribe the avoidance of losses with certainty (Fig. 1). For intermediate levels of T, individual and collective interests diverge as the NE prescribes progressively lower contributions, while the CS prescribes full loss avoidance. For T=2100, the threshold used in our experiment, the NE prescribes no contribution – regardless of the individual's degree of risk aversion (Point A) - while the CS for risk-neutral agents prescribes a PLA of 69% (Point B). If agents are risk-averse, the CS prescribes a higher PLA (Point C) than for risk-neutral agents, which is in general lower than certain loss avoidance (Point D). Accordingly, the interaction implemented in our experiment had the typical characteristics of a social dilemma27 with individual interests maximized by no contribution to the collective account and group interests maximized by positive contributions.

Results

Cooperation increases in international interaction with sanctions

In stark contrast to the NE prediction, all groups achieved substantial levels of loss avoidance. The PLA ranged from 12% to 100%, the grand mean being 70.1%, in line with the CS prediction for risk-neutral agents (Fig. 2). Neither contributions nor sanctions differed significantly between the two locations within each country (Tables S7-S8). Accordingly, we consider aggregate national observations only. We first analyze the S-treatments.

86% of groups in the S-treatments achieved a PLA higher than the CS, with seven groups achieving full loss avoidance. In national treatments with sanctions, German groups achieved significantly higher PLA than Russian groups (d=1.39; p=0.0005; N=32; d is Cohen's d; all tests are two-sided Wilcoxon-Mann-Whitney (WMW) tests unless otherwise indicated), confirming previous comparative research[18-19]. International S-treatments achieved similar levels of cooperation to national German S-treatments, with no significant differences in the distribution of PLA (Kruskal-Wallis test (KW); p= 0.41; N=48). By contrast, the average PLA (APLA) in the Russian national S-treatment was significantly lower than both the International B-treatment (d= -1.41; p=0.0005; N=32) and O-treatment (d= -1.18; p=0.0037; N=32).

Cooperation in international S-treatments was thus significantly higher than in the Russian national S-treatment. This may be due to German participants increasing contributions in international S-treatments to compensate for Russian participants' lower cooperation rates. An alternative is Russian participants increasing their cooperation in international S-treatments as opposed to national treatments. The latter

alternative is the correct one. There is no significant difference between German and Russian cooperation in either the B-treatment (d=-0.016; p=1.00; N=16) or the O-treatment (d=0.11; p=0.84; N=16), using two-tailed WMW matched-pairs signrank tests (Fig. 3, Panel A).

In the initial periods, contributions by Russian participants in international S-treatments started below German participants' cooperation but quickly caught up as interactions continued (Fig. S5, Panels A-B). Non-parametric tests reveal that while in periods 1 and 2 Russian participants' contributions in international treatments were not significantly different from Russian participants' contributions in the national S-treatment, contributions by Russian participants were significantly higher in international treatments than in national treatments in all subsequent periods (Table S9). This is significant at least at the 1% level (or less) in the B-treatment and at least at the 5% level (from round 4) in the O-treatment. Conversely, the hypothesis of equality of distributions for contributions in international and national treatments is never rejected for German participants in any period. We can thus conclude that in international treatments, Russian participants' contributions quickly increased in comparison with the national treatment and converged to German participants' contributions. This result suggests that international cooperation with sanctions was beneficial overall because Russian participants achieved higher PLA while PLA remained stable for German participants. Moreover, there was no significant difference in payoffs accruing to German and Russian participants (d=-0.17; p=0.73, N=64).

International cooperation increases less without sanctions

Can the same benefits be achieved without sanctions? Overall, the PLA was 18% lower in NS-treatments than S-treatments (d=-1.02; p<0.0001; N=128; Fig. 1). Only 34% of groups in NS-treatments exceeded the PLA prescribed by the CS. S-treatments yielded significantly higher PLA than NS-treatments in German national treatments (p=0.002; N=32), in international B-treatments (d=-1.25; p= 0.0033; N=32), and in international O-treatments (d=-1.07; p= 0.0018; N=32). However, this was not the case in Russian national treatments at conventional levels of significance, (d=-0.64; p=0.07; N=32). Without sanctions, German groups did not achieve significantly higher PLA than Russian groups, again at conventional levels (d=0.70; p=0.08; N=32).

As in the S-treatments, there is no significant difference in PLA between the two international NS-treatment and the German national NS-treatments (KW Test: p=0.99; N=48). However, PLA in international NS-treatments is here not significantly higher at conventional levels than in the Russian national NS-treatment (d= -0.71; p= 0.062, N=32 for INT_B_NS; d=-0 .72; p= 0.055, N=32, for INT_O_NS). Cooperation by Russian participants in international treatments is again not significantly different from German participants' cooperation (d=-0.03; signrank WMW: p=0.88, N=16 in INT_B_NS; d=0.18; signrank WMW: p=0.64 in INT_O_NS; Fig. 3B), nor is it significantly higher, at conventional levels, than cooperation in the national NS-treatment (d= -0.66; p= 0.067; N=48). Unlike in S-treatments, Russian participants generally did not significantly increase their contributions in international NS-treatments compared to the national NS-treatment in individual periods of interaction (Table S9). We decompose the treatment effects of cooperation by Russian and German participants in Table S10 and Fig. S6, pooling the two international

treatments. Introducing sanctions in national interactions increases cooperation by 13% in comparison with the national NS-treatment. Remarkably, the same increase is effected without sanctions by "internationalizing" interaction – i.e., having Russians interact with Germans. While neither of these effects is statistically significant, introducing sanctions in an international context increases Russian participants' cooperation by 20% in comparison to either the Russian national S-treatment or the international NS-treatment. As we have noted, both these increases are statistically significant. We can thus conclude that, while sanctions alone and internationalization alone brought about only marginal increases in cooperation by Russian participants, the combination of the two factors was necessary to significantly increase Russian participants' cooperation.

Low sanctioning suffices to spur cooperation

Next, we analyze the mechanisms that made sanctions effective in increasing cooperation. Only 7% of the available endowment was spent on sanctions and in 70% of cases no sanction was administered (Fig. S7). Russians spent about 68% more than Germans in national treatments (Fig. S8), but the difference is not significant (d=-0.57; p=0.16, N=32). Germans used sanctions significantly more often in the O-treatment than in the national S-treatment (d= -0.70; p=0.021, N=32) and more than in the B-treatment – though not significantly (d= -0.59; p=0.052, N=32). We decompose sanctioning into pro-social (*PS*) and anti-social (*AS*) sanctioning (SI: Section S1.4). We define *AS* as instances in which an *ego* punished an *alter* who contributed no less than the group median, while *PS* is the residual category[26]. *AS* is puzzling because it targets individuals who are increasing social welfare in the group, but it has proved to be endemic in experiments with people from cultural areas classified as orthodox/post-communist[19]. In national treatments, Russians spent 2.52 times as much as Germans on *AS,* although the difference is not significant at conventional levels (d= -0.84; p=0.055, N=32; Fig. S9). Germans significantly increased *PS* in the International O-treatment compared to the German national treatment (d=0.60; p=0.044, N=32), presumably because they were initially faced with more low cooperators than in the national treatment. Russian participants tended to lower *AS* and increase *PS* in international as opposed to national treatments, but these changes are not significant (SI: Section S1.4). We also note that sanctions had a spike in the last period when no counter-sanctioning could have occurred (Fig. S10). This spike can only be accounted for as revenge for previous interactions, as it could not have any disciplinary function.

Previous research has found that sanctions are effective if people increase cooperation after having been sanctioned[19]. With an OLS econometric model controlling for period effects, we compute the impact on the contribution made in the next period of having a token deducted through sanctioning. On average, a token deducted by sanctioning raised cooperation by 0.42 tokens in the next period (p<0.001; Table 2, Column 1), but the effect differed significantly across treatments (Table 2, columns 2-5). The effectiveness of sanctioning was highest in the International B-treatment, where a token deducted by sanctions increased cooperation by 0.72 tokens, and lowest in the International O-treatment, where sanction effectiveness was less than half that amount. Hence, sanctions lost part of their effectiveness when nationality was revealed to participants than when it was concealed from them. In the SI: S1.5, we show that these results are robust to the introduction of demographic controls and analyze the effect of

being sanctioned regardless of the amount of sanctioning (Table S12). Among demographic characteristics, we find that men contribute significantly less, sanction significantly more, and are significantly less responsive to sanctions than women (SI: Sections S1.5-S1.7; Tables S11-S14).

Payoffs are higher in no-sanction treatments.

Finally, we analyzed the effects of different treatments on final payoffs. Average expected individual payoffs were significantly higher in NS-treatments (263.9 tokens) than in S-treatments (246.6 tokens) ($p<0.0001$; N=128). This is also the case in every pairwise comparison of NS- and S-treatments in either National or International treatments (Fig. S11 and SI: section S1.8). By construction, the CS for risk-neutral agents maximizes expected group payoffs. This implies that, as participants in the S-treatments contributed more than what is prescribed by the CS, they incurred a cost in comparison with the optimal contribution level. An interpretation of this result is that individuals are predominantly risk-averse and will thus collectively prefer a level of APLA such as Point C in Fig. 1. Nonetheless, such a high level of APLA can only be achieved when sanctions are available. Without sanctions, it is plausible that, as posited by previous research[19], participants will withhold cooperation as a form of indirect punishment, thus lowering the APLA.

Discussion

Many fear that as global-level cultural heterogeneity, complexity, and institutional limitations make international cooperation even more difficult than the local or national variety[27-29], international cooperation will be unable to steer clear of a tragedy of the "global commons"[30-31]. Our results offer a glimmer of hope, indicating that the combination of sanctions and the internationalization of interaction brings about net positive effects. A plausible interpretation of this is that after observing higher-than-expected cooperation in the initial periods of interaction, Russian participants involved in international interactions were quick to revise their beliefs on their counterparts' willingness to cooperate. Arguably, Russian participants' initial beliefs were rooted in the cooperation rates observed in the environment with which they were most familiar, i.e. local or national interaction. Consistently with a motivational model of conditional reciprocity[18,21], adjusting beliefs upwards prompted Russian participants to be more cooperative in international than in national interactions. Sanctions were however still necessary to achieve this result. Our study confirms that in national contexts the effectiveness of sanctions is culture-specific[13,19] and further shows that German high cooperators were as capable of disciplining low cooperators in international interactions as in national ones. The substantial similarity of results in Blind and Open treatments show that it was the actual content of participants' actions, rather than motivations linked to awareness of the specific nationalities involved, that determined the beneficial effects of international cooperation.

Though our experiment reproduces in a stylized fashion various features of the consequences of climate change and trade sanctions on individual earnings, the problem of "scalability" is apparent in connection with the outcome of our experiment[31-32]. Nonetheless, at a more fundamental level, our experiment can

be seen as revealing the willingness of the general population to abide by an agreement once an agreement has been reached[34], which is a fundamental feature of any international agreement[2]. In fact, individuals who cooperated in the experiment were also more likely to conduct environmentally sustainable behavior in real life, such as buying environmentally-friendly goods, saving water, participating in ecological movements, and recycling. Experimental contribution is positively associated with an index of such environmentally sustainable actions (b= 3.11; p=0.18; N=744; Fig. S12-S13), the relationship being at the margins of significance in S-treatments (b=4.84; p=0.085; N=372; SI: Section 1.5-1.7; Table S14). This offers some evidence, albeit weak, that behavior in our experiment is linked to willingness to take action to preserve the environment in real life. It is an open question whether our results would extend to other types of public-good problems, as it has been argued that cooperation tends to be higher when the public good consists of increased protection against collective risk, as in our experiment, rather than positive externalities on others' payoffs[35].

Despite these limitations, our findings suggest that sanctions can be used in international interactions to discipline people who would otherwise not cooperate and that they can do this without risking a spiral of retaliation and counter-retaliation. This evidence supports the view that international sanctions can lead to significant and robust changes in standards of conduct and should be used more extensively in international agreements, particularly in climate agreements. Our study has also shown a preference for remarkably high levels of collective loss-avoidance, and such preferences should be addressed by policy-makers.

Methods

Experiments in the social sciences

For several decades the social sciences have applied experimental methods to the study of social interaction in controlled laboratory settings. Our study was characterized by features typical of experiments: (a) Monetary incentivization: Participants received money endowments from the researchers and made decisions on how to allocate these endowments. Participants were then paid the monetary payoffs resulting from their choices and the other group members' choices. Mean earnings were 25.00€ in Germany and 750 Ruble in Russia. (b) Anonymity: Participants' real identity was not revealed to other participants in the course of the experiment. Participants were instead identified through a randomly assigned numeric label so that other participants could reconstruct the "history" of other participants' actions. Reputation-building and revenge were then feasible motivations in our experiment. (c) Lack of deception: Participants were never deceived on any aspect of our design. (d) Treatment randomization: Randomization of treatments (see Table 1) occurred at the session-level (SI: section S4.4). Randomization permits causal inference from the treatments to the main variable of interest – namely, cooperation.

Sampling university students is subject to three types of biases: (a) A self-selection bias concerning the university student population; (b) A bias caused by participants performing more socially desirable

behavior when interacting in the lab than in real-life; (c) A lack of representativeness of university student vis-à-vis the general population. We discuss such biases extensively in SI: Section S1.9. Based on specifically designed experimental studies, we conclude that (a) the bias between university students who self-select into experimental studies and the full population of university students appear to be negligible[36]. (b) Even if some studies show more socially desirable behavior in experiments than in real life[37-38], other studies do not detect this effect[36]. Most importantly, experiments permit tightly controlled variation in the main parameters of the interaction, thus enabling causal inference. This would be in most cases impossible in natural settings[39]. (c) Even if several studies find that adult samples behave more pro-socially than students' sample[40-41], this does not prevent causal inference as long as treatment effects are also distorted by the type of population being sampled. It is then reassuring that correlations across a broad range of variables have similar size in university students' samples and samples representative of the general population[36]. In fact, less noise and fewer cognitive errors have been found in student samples than in representative adult samples[36], which suggest that students' sample may be more reliable than representative samples for hypotheses testing. To further test the generalizability of our results, we conducted an out-of-sample estimation to simulate the size of the treatment effects had we used representative samples of the populations. The result of this exercise, reported in SI: Section S1.9 and Table S15, is that the main treatment effects – in particular, the higher cooperation rate in international treatments with sanctions compared to national treatments – would hold even with nationally representative samples.

Measures to ensure between-country comparability of data

We followed best practices in cross-country comparative research[42-45] to address some of the aspects that can compromise cross-country data comparability. We sought to minimize experimenter effects by having lead researchers conducting the sessions following the same experimenter script (available at https://osf.io/k4d8w/). Several aspects of the experiment protocol, in particular instructions and summaries (SI: section S5 for instructions and https://osf.io/ch4gd/ for Powerpoint© summaries), were also standardized. We run a pilot session in which all the four lead experimenters were present with the goal of unifying the style of conduction of the four lead experimenters. Language effects were minimized using the technique of back-translation. Currency effects were controlled by referring to 'tokens' rather than to national monetary units, and by ensuring that the token value in the two countries had equivalent purchasing power (see SI: Section 4.1 for details).

The experiment protocol, instructions and questionnaire are reported in SI: section 4. Experiment script guiding researchers in the conduction of the research session, and the software to visualize instructions and collect data are available at a repository of the Open Science Foundation (https://osf.io/4s6p3/). The experiment protocol has been deposited at Protocol Exchange: https://doi.org/10.21203/rs.3.pex-1459/v1.

Determination of sample size

We anchored the sample size in our study to the sample size of other studies with a similar design to ours[16,23]. In these studies, the unit of observation is a group of participants, and each group comprises 6-10 participants (we chose the lower bound of 6 for our experiment). These studies had 10 groups per treatment and found a very large effect size for their treatments. In particular, the size of the effect of introducing uncertainty over the safety threshold in one of these studies[23] was Cohen's d=3.59 {m1 = 150.9, m2 = 79.9, sd1 = 7.69, sd2 = 26.90}. We were skeptical that in the context of our study, in which the main treatment concerns the variation in cooperation in an international environment vis-à-vis a national one, the effect size would have been as large. Therefore, we decided to increase sample size to N=16 per treatment. Ex post power analysis confirmed that our prediction was correct. The sample size requested for Type-1 error = 0.05 ad for Power = 0.80 to detect a significant difference in the means observed in one of our key treatments (the difference of cooperation in the International Open treatment and the National Russian treatment under sanctions, where {m1 = 22.4375, m2 = 28.7875, sd1 = 6.9067, sd2 = 4.6133}) is N=15, which is very close to our choice of N=16. The size of this effect is Cohen's d=-1.16.

Ethical approval and data protection

Since our research could not provide any harm to participants and did not involve any medical treatment, the approval by an ethics committee or institutional review board was waived by our universities. We asked every participant to read an information sheet and sign an informed consent form before starting the research session. (see SI: section S4.3 for further details and for measures to ensure full data anonymization).

Statistical methods

As reported in the main text, we used two-sided non-parametric tests for our tests, taking group outcomes as the unit of observations. We used Cohen's d as the measure for effect size. Econometric models have been fitted to examine possible determinants of individual behavior. Such models are described in SI: Section S1.5-S1.7.

Data Availability Statement

The dataset generated and analyzed during the current study are available in the project repository of the Open Science Foundation: https://osf.io/r3a2x/?view_only=97916d398f1745b2b0e7ad862f4ecb6e . Analyses codes will be made available upon reviewers' request and before the publication of the paper in our project repository.

Declarations

Acknowledgments

We acknowledge funding from the Center for Global Cooperation Research at the University of Duisburg-Essen, Christian-Albrechts-University of Kiel, and the National Research University Higher School of Economics (HSE) - within the framework of a subsidy by the Russian Academic Excellence Project ’5-100. We are especially grateful to all research assistants involved in this Project, who are listed at SI: Section 8.

Authors’ contribution: All authors conceived research and collected data. AB, HHS, TR, MR processed data. AB, GG and HHS analyzed data. TR developed theoretical analysis. GG drafted manuscript. All authors reviewed and commented on manuscript.
Competing Interest Statement: Authors declare no competing interests.

References

1. W. Nordhaus, Climate clubs: Overcoming free-riding in international climate policy. *Am. Econ. Rev.* **105**, 1339-70 (2015).

2. S. Barrett, Why cooperate? The incentive to supply global public goods (Oxford University Press, 2007).

3. R. Aichele, G. Felbermayr, Kyoto and the carbon footprint of nations. *J. Environ. Econ. Manag.* **63**, 336-354 (2012).

4. M. Nowak, and Karl Sigmund. "Evolution of indirect reciprocity." *Nature* 437.7063 1291-1298 (2005).

5. C. Handley, and S. Mathew. Human large-scale cooperation as a product of competition between cultural groups. *Nature Comms*, **11**, 1-9 (2020).

6. Mathew, S., & Boyd, R. Punishment sustains large-scale cooperation in prestate warfare. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 11375-11380 (2011).

7. J. Choi, S. Bowles, The coevolution of parochial altruism and war. *Science* **318**, 636-640 (2007).

8. H., Bernhard, U. Fischbacher, & Fehr, E. Parochial altruism in humans. *Nature* **442** (7105), 912-915 (2006).

9. A. Romano, M. Sutter, J. H. Liu, T. Yamagishi, & D. Balliet. National parochialism is ubiquitous across 42 nations around the world. *Nature Comms,* **12**, 1-8 (2021).

10. A.R. Dorrough, A. Glöckner. Multinational investigation of cross-societal cooperation. *P. Natl. Acad. Sci. U.S.A.*, **113**, 10836-10841 (2016).

11. R. Keohane, D. Victor. Cooperation and discord in global climate policy. *Nat. Clim. Change* **6**, 570–575 (2016).

12. W. Przepiorka, & A. Diekmann. Individual heterogeneity and costly punishment: a volunteer's dilemma. *Proc. Royal Soc. B,* **280** (1759), 20130247 (2013).

13. J. Henrich, R. McElreath, A. Barr, J. Ensminger, C. Barrett, A. Bolyanatz, ... C. Lesorogol. Costly punishment across human societies. *Science*, **312**, 1767-1770 (2006).

14. R. Boyd, H. Gintis, S. Bowles, P.J. Richerson. The evolution of altruistic punishment. *P. Natl. Acad. Sci. U.S.A.*, **100**, 3531–3535 (2003).

15. A. Berger, C. Brandi, D. Bruhn. *Environmental provisions in trade agreements: promises at the trade and environment interface.* (German Development Institute, Briefing Paper, ***16***, 2017)

16. M. Milinski, R. Sommerfeld, H.-J. Krambeck, F. Reed, J. Marotzke. The collective-risk social dilemma and the prevention of simulated dangerous climate change. *P. Natl. Acad. Sci. U.S.A.* **105**, 2291-2294 (2008).

17. M. Alexander, F. Christia. Context modularity of human altruism. *Science*, **334**, 1392-1394 (2011).

18. S. Gächter, B. Herrmann. Reciprocity, culture and human cooperation: previous insights and a new cross-cultural experiment. *Phil. Trans. R. Soc. B*, **364**, 791–806 (2009).

19. B. Herrmann, S. Gächter, C. Thöni. Culture and cooperation. *Phil. Trans. R. Soc. B*, **365**, 2651-2661 (2010).

20. F. Gino, S. Ayal, D. Ariely. Contagion and differentiation in unethical behavior: The effect of one bad apple on the barrel. *Psychol. Sci.*, **20**, 393-398 (2009).

21. U. Fischbacher, S. Gächter, & E. Fehr. Are people conditionally cooperative? Evidence from a public goods experiment. *Econ. Lett*. **71**, 397-404, (2001).

22. E. Kriegler, J.W. Hall, H. Held, R. Dawson, H.J. Schellnhuber. Imprecise probability assessment of tipping points in the climate system. *P. Natl. Acad. Sci. U.S.A.*, **106**, 5041-5046 (2009).

23. S. Barrett, A. Dannenberg. Climate negotiations under scientific uncertainty. *P. Natl. Acad. Sci. U.S.A.*, **109**, 17372-17376 (2012).

24. M. Finocchiaro Castro. Where are you from? Cultural differences in public good experiments. *J. Socio-Econ.* **37**, 2319-2329 (2008).

25. R. Böhm, B. Rockenbach. The inter-group comparison–intra-group cooperation hypothesis: comparisons between groups increase efficiency in public goods provision. *PLOS ONE*, **8**, e56152 (2013).

26. N. Nikiforakis. Punishment and counter-punishment in public good games: can we really govern ourselves? *J. Publ. Econ*, **92**, 91–112 (2008).

27. E. Ostrom, J. Burger, C.B. Field, R.B. Norgaard, D. Policansky. Revisiting the commons: local lessons, global challenges. *Science*, **284**, 278-282 (1999).

28. Schmidt, K. M., & Ockenfels, A. Focusing climate negotiations on a uniform common commitment can promote cooperation. *P. Natl. Acad. Sci. U.S.A.*, **118** (2021).

29. A.Tavoni, A. Dannenberg, G. Kallis, A. Löschel. Inequality, communication, and the avoidance of disastrous climate change in a public goods game. *P. Natl. Acad. Sci. U.S.A.*, **108**, 11825-11829 (2011).

30. G. Hardin. The tragedy of the commons. *Science* ***162***, 1243-1248 (1968).

31. J. Jacquet, K. Hagel, C. Hauert, J. Marotzke, T. Röhl, M. Milinski. Intra-and intergenerational discounting in the climate game. *Nat. Clim. Change*, **3**(12), 1025-1028 (2013).

32. O. Al-Ubaydli, J.A. List, D.L. Suskind. What can we learn from experiments? Understanding the threats to the scalability of experimental results. *Am. Econ Rev.*, **107**, 282-86 (2017).

33. D. Messner, A. Guarín, D. Haun, "The behavioral dimensions of international cooperation" in D. Messner, S. Weinlich, Eds. *Global Cooperation and the Human Factor in International Relations* (Routledge, Abingdon, 2015), pp. 65-83.

34. F. Guala, Reciprocity: weak or strong? What punishment experiments do (and do not) demonstrate. *Behav. Brain Sci.* **35**, 1-15 (2012).

35. A Ispano, P. Schwardmann. Cooperating over losses and competing over gains: A social dilemma experiment, *Game Econ Behav.* **105**, 329-348 (2017).

36. E. Snowberg, L. Yariv. Testing the waters: Behavior across participant pools. *Am. Econ Rev.* **111**, 687-719 (2021).

37. S. D. Levitt, J.A. List. What do laboratory experiments measuring social preferences reveal about the real world?. *J. Econ. Perspect.* **21**, 153-174 (2007).

38. M.M. Galizzi, D. Navarro-Martinez. On the external validity of social preference games: a systematic lab-field study. *Manage. Sci.*, **65**, 976-1002 (2019).

39. A. Falk, J.J. Heckman. Lab experiments are a major source of knowledge in social sciences. *Science*, **326**, 535-538 (2009).

40. A.W. Cappelen, K. Nygaard, E. Ø. Sørensen, B. Tungodden, B. Social preferences in the lab: A comparison of students and a representative population. *Scand. J. Econ.* **117**, 1306-1326 (2015).

41. A. Falk, S. Meier, C. Zehnder. Do lab experiments misrepresent social preferences? The case of self-selected student samples. *J. Eur. Econ. Assoc.* **11**, 839-852 (2013).

42. A. E. Roth, V. Prasnikar, M. Okuno-Fujiwara, S. Zamir. Bargaining and market behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An experimental study. *Am. Econ Rev.* **81**, 1068-1095 (1991)

43. Herrmann, B., Thöni, C., Gächter, S. Antisocial punishment across societies. *Science* **319**, 1362-1367 (2008).

44. N.R. Buchan, G. Grimalda, R. Wilson, M. Brewer, M., E. Fatas, M. Foddy. Globalization and human cooperation. *P. Natl. Acad. Sci. U.S.A.* **106**, 4138-4142 (2009).

45. S.J. Goerg, H. Hennig-Schmidt, G. Walkowitz, E. Winter. In wrong anticipation – miscalibrated beliefs between Germans, Israelis, and Palestinians. *PLOS ONE*, **11**, e0156998 (2016).

Tables

Due to technical limitations, the tables are only available as a download in the supplementary files.

Figures

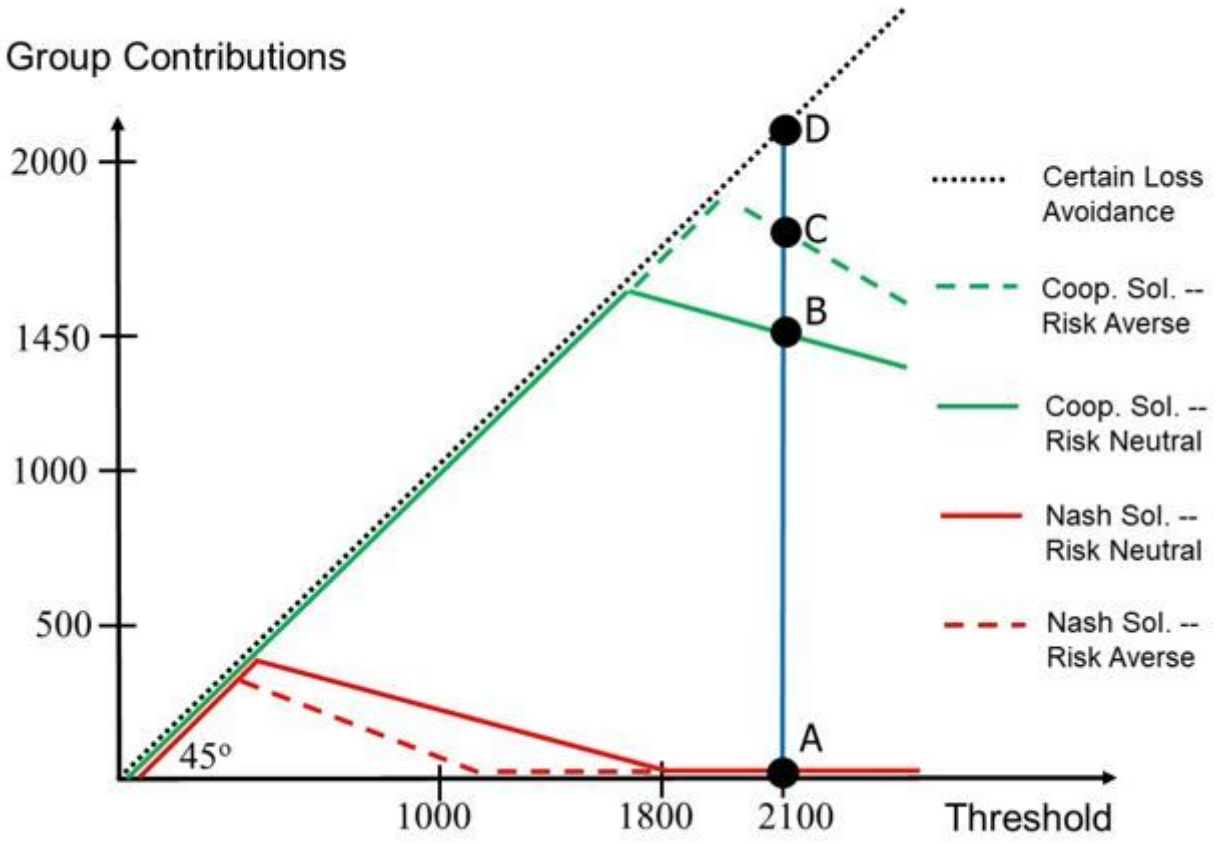

Figure 1

Nash Equilibria and Cooperative Solutions in the CRSD game for different levels of the certain safety threshold T (x-axis). The y-axis plots the group-level contribution for each solution.

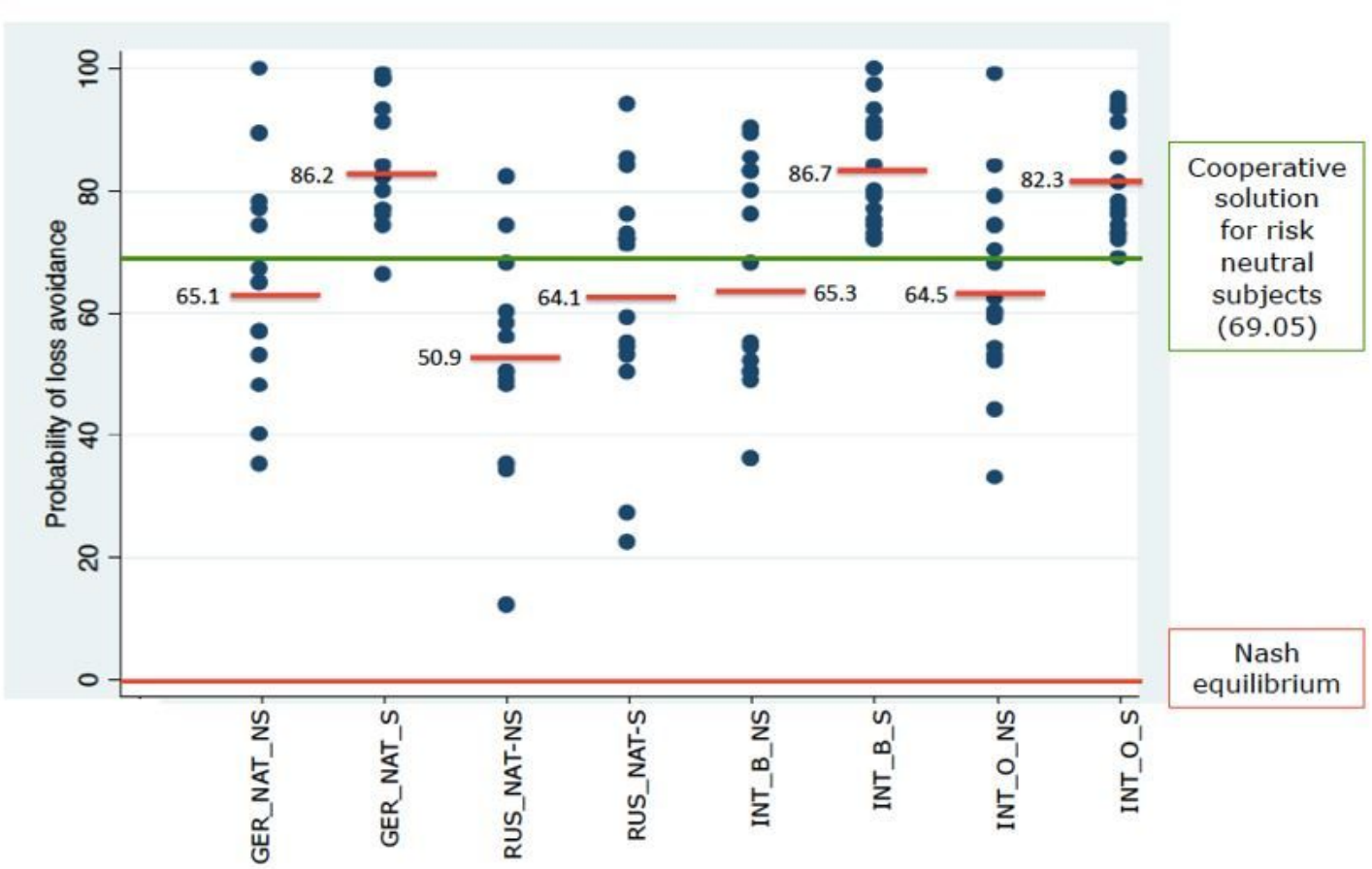

Figure 2

Probability of loss-avoidance for each group and treatment.

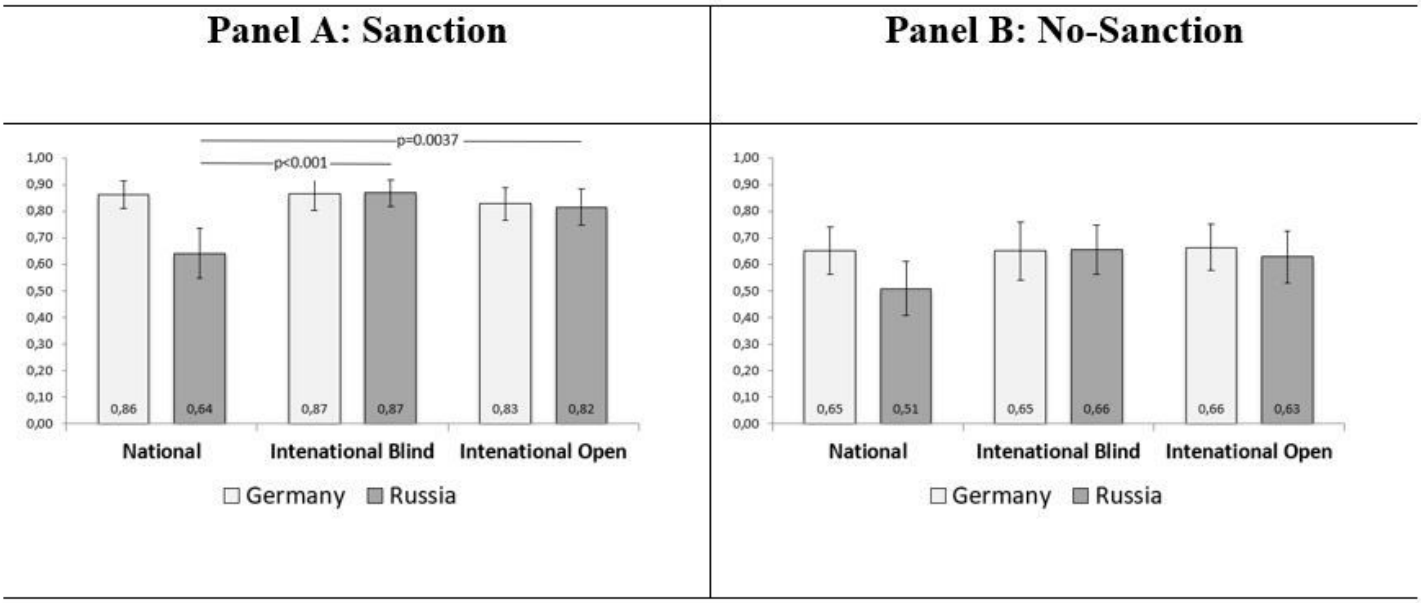

Figure 3

Average cooperation rates by nationality and treatment. Mean contributions to the collective fund as percentage of the level needed to avoid loss with certainty. Error bars are 95% confidence intervals with bootstrapped standard errors (10,000 repetitions).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryInformation.pdf
- tables.docx