Article:

Del Ponte, A., Masiliūnas, A. and Lim, N. (2025) Decentralized voluntary agreements do not reduce emissions in a climate change experiment. Ecological Economics, 227. 108438. ISSN: 0921-8009

https://doi.org/10.1016/j.ecolecon.2024.108438

Decentralized voluntary agreements do not reduce emissions in a climate change experiment

Abstract

Can climate accords based on decentralized and voluntary agreements successfully reduce carbon emissions? We designed an economic experiment to study the effectiveness of the best-known mechanisms to foster international cooperation on climate change mitigation: climate pledges, financial penalties, and peer evaluation. We test each mechanism both separately and together. In the climate pledge treatments, participants could pledge their desired emissions target, approved by majority vote. In the treatments with financial penalties, failure to meet pledges triggered monetary sanctions. In the peer evaluation treatments, participants could evaluate each other, which determined who would receive an additional nonmonetary environmental prize. We find that most participants joined climate agreements and met their pledges, but pledges were insufficiently ambitious. As a result, neither pledges, financial penalties, nor peer evaluation reduced emissions. These results question the effectiveness of decentralized and voluntary climate agreements, such as the Paris Agreement.

Introduction

Climate change is a global social dilemma as many countries are unwilling to sufficiently reduce carbon emissions. To solve this collective action problem, the international community relies on climate treaties, which should provide compelling incentives to sufficiently reduce emissions (Barrett, 2003). Unfortunately, both the Kyoto Protocol and the Paris Agreement have so far been unsuccessful, as countries are behind schedule in meeting their targets (du Pont & Meinshausen, 2018; Roelfsema et al., 2020) and global warming will likely exceed 1.5°C during the 21st century even if the commitments made in 2021 are achieved (IPCC, 2023). Our study uses theory and experiments to explore the potential reasons of this failure and test whether decentralized and voluntary climate agreements can be successful.

We introduce a novel public bads game to compare the effectiveness of a pledge-and-review procedure without any enforcement mechanism against the procedures with monetary penalties or non-monetary peer evaluation. We compare various combinations of these mechanisms using a 2 (Peer evaluation / No peer evaluation) x 3 (No pledges / Pledge-and-review / Pledge-and-review with penalties) between-subjects design. We chose these three features because they are key components of actual climate treaties. The Paris Agreement relies on pledge-and-review with peer evaluation, as countries periodically make and update their nationally determined contributions (NDCs), and those who fail to comply or make unambitious pledges are informally sanctioned (Jacquet & Jamieson, 2016). Financial penalties for exceeding the target amount were used in the Kyoto Protocol in the form of a market for assigned amount units[1]. Various elements of each treaty have been compared and evaluated (Barrett, 2003; Falkner, 2016; Held & Roger, 2018; Young, 2016), but it is unclear if

[1] The Kyoto Protocol included compliance procedures, detailed in the Marrakesh Accords, authorizing the Enforcement Branch to apply punitive measures to the countries that fail to comply (Halvorssen & Hovi, 2006; Hovi et al., 2007; Nentjes & Klaassen, 2004). In practice, the compliance system had a number of weaknesses, which could allow non-compliant countries to avoid penalties by postponing them indefinitely or withdrawing from the agreement (Barrett, 2003).

the treaties would have been more successful with a different design. We address this problem by running controlled laboratory experiments, where the design elements can be systematically varied.

By comparing a treatment with penalties to treatments without penalties and a baseline without agreements, we can separately identify the effectiveness of agreements and monetary penalties. Overall, previous studies uncovered various mechanisms to increase the provision of public goods (see surveys in Chaudhuri, 2011 and Ledyard, 1995). Unfortunately, many of these solutions would be difficult to implement in climate change negotiations and thus have limited use for designing climate treaties. For example, peer punishment tends to increase contributions (Fehr & Gächter, 2000), but in practice, states would find it difficult to impose punishment, such as trade sanctions, on countries that failed to meet their emissions targets.

Our design is novel in several ways. We are the first to test the effectiveness of automatic financial penalties for exceeding the emissions target. In previous literature, the agreements were either non-binding (Feige et al., 2018; Tavoni et al., 2011), punishment was allocated at the discretion of the other participants (Dannenberg, 2016), or compliance was automatically enforced, either in full (Kosfeld et al., 2009; Lippert & Tremewan, 2021; Schmidt & Ockenfels, 2021) or partially (Chávez et al., 2023). Instead, we introduce automatic financial penalties for *each unit* produced above the pledge, following the design of the Kyoto Protocol. Second, we study how peer evaluation interacts with financial penalties, allowing us to identify whether peer evaluation, alone or in combination with financial penalties, can nudge countries to reduce their emissions. We also contribute to the recent literature that experimentally evaluates the features of the Paris Agreement, such as increasing commitments over time using a ratchet-up mechanism (Alt et al., 2023; Cherry et al., 2021; Gallier & Sturm, 2021).

Literature Review

We study the effectiveness of pledge-and-review mechanisms in a public bads game designed to mimic the incentives that countries face when mitigating climate change. While we are not aware of previous literature that studied the effectiveness of climate agreements in this setup, similar interventions have been examined in other types of social dilemma games, such as the voluntary contribution mechanism (VCM). A key finding from this strand of the literature is that non-binding pledges fail to reduce contributions. In these experiments, participants can typically enter non-binding agreements before deciding how much to contribute to the public good. In Dannenberg (2016), the common commitment was set exogenously and the agreement was formed if all participants joined. Afterwards, participants chose their contribution in the VCM without being bound by the agreement. The study found that the agreement institution did not increase contributions. In Feige et al. (2018), each participant proposed a vector of individual contributions and voted for each proposal. A proposal was accepted if all participants voted for it. The study also found that this institution did not increase contributions in a threshold public goods game. In our experiment, participants pledge not to exceed a certain amount, but then the group votes on whether to accept these pledges, modeling group negotiation. In contrast, the studies in which agreements require group consensus have pledges that are either exogenous (Dannenberg, 2016), or determined by one of the group members (Feige et al., 2018). Our procedure also had an additional stage where each participant decided whether or not to join the agreement, modeling the ratification of climate treaties.

Also closely related is a paper by Schmidt & Ockenfels (2021), who compared an “individual commitment” treatment where participants submitted their own pledges to two other treatments where participants proposed either a set of commitments for each group member (“complex common commitment”) or a single commitment that applied to all group members (“uniform common commitment”). The study found that uniform common commitments increased contributions, although these treatments were not compared to a baseline with no commitments. Schmidt & Ockenfels (2021) also differs from our study in other ways. For instance, we use a symmetric public bads game in which pledges are accepted by majority voting, whereas Schmidt & Ockenfels (2021) used an asymmetric VCM where pledges were accepted only if all participants agreed. Moreover, we study financial penalties for non-compliance, whereas the agreements in Schmidt & Ockenfels (2021) are either non-binding or fully enforced.

Less related are the studies where participants can select their own pledges but the group cannot negotiate or vote over them. McEvoy et al. (2022) found that non-binding pledges were ineffective in a mitigation-only scenario, and only temporarily effective when adaptation was also an option. Several other experiments found that such one-way communication of intentions did not affect average contributions in the VCM (Bochet & Putterman, 2009), the symmetric threshold public goods game (Tavoni et al., 2011), and the two-stage public goods game (Cherry et al., 2021). In contrast, there is some evidence that free-form communication improves cooperation in social dilemmas (Balliet, 2010).

Another strand of literature studied whether cooperation is facilitated by non-monetary sanctions through peer evaluation. Barrett & Dannenberg (2016) tested how informal sanctions affect pledges and contributions in a modified threshold public goods game. In the experiment, participants first proposed a group target, then made individual pledges, and finally chose their contribution. Depending on the treatment, participants could grade their peers at various stages of the game, but such peer review did not have a significant effect on contributions. Other studies found mixed evidence about the effects of non-monetary punishment on contributions in the VCM. Masclet et al. (2003) found that non-monetary punishment increased contributions for a short time, but the effectiveness soon waned, in contrast to the more persistent effect of monetary punishment. A positive yet transient increase in contributions was also documented by Faillo et al. (2020), who also ran a control treatment with no penalties. Dugar (2013) found that the ability to send disapproval points increased contributions, but the improvement was even higher when participants could express both approval and disapproval; the treatment with only approval points was no different from the baseline. In contrast, Peeters & Vorsatz (2013) found that an option to signal feedback with emoticons did not have a significant effect on contributions. Overall, the literature suggests that the effect of peer evaluation ranges from mildly positive to none, dissipates over time, and is weaker compared to pecuniary sanctions.

More generally, our study contributes to the experimental research on endogenous institution formation in social dilemma games (see a survey in Dannenberg and Gallier, 2020). Closest to our treatment with financial penalties are the studies where the group votes on introducing an institution that automatically sanctions participants who do not fully contribute to the public good (Gallier, 2020; Tyran & Feld, 2006; Vollan et al., 2017). Also related is a study by Cherry and McEvoy (2013), where the institution is created if all participants join the agreement (or if the majority join, in another treatment) and sanctions are imposed by withholding previously transferred deposits. When sanctions are sufficiently high to theoretically deter free-riding, experiments find a significant increase in contributions (treatment with a high penalty in Tyran and Feld, 2006, and the treatment that mandates full participation in Cherry and McEvoy, 2013). However, when the sanctions are lower, the results are mixed: Tyran and Feld (2006) and Vollan et al. (2017) find no significant treatment effects, but Gallier (2020) finds a significant increase in contributions.

Our study differs from the literature on endogenous institution formation in several aspects. First, our institution more closely resembles the sanctions that have been used in climate change agreements, such as the Kyoto protocol. Instead of imposing a fixed penalty on the participants who fail to fully contribute, participants pay a penalty for each unit produced above their pledge. Second, sanctions in our institution depend on participants' own pledges, which are a key element in the real-life pledge-and-review process. Third, in previous literature, all participants were bound by the institution if the group voted to introduce it. In contrast, participants in our experiment can select whether to join the agreement once it has been formed. This additional decision reflects the voluntary nature of climate treaties, which need to be nationally ratified. Overall, our institution is decentralized (participants select their own pledges) and voluntary (participants choose whether to join the agreement), which was not the case in the previously studied institutions. Finally, the previously studied institutions, if introduced, necessarily changed the incentives for all the participants, even if the sanctions were not sufficiently high to completely deter free-riding. In contrast, participants in our experiment can avoid all penalties by choosing unambitious pledges or not joining the agreement.

Experimental Design

Our design models decentralized and voluntary agreements in international cooperation on climate change mitigation, such as the Paris Agreement. Without a governing international organization, it is up to the countries to negotiate the terms of the agreement, and no country can be forced to enter the agreement or abide by a certain emissions target. We model the voluntary nature of agreements by allowing the participants in pledge-and-review treatments to propose their own targets, vote on whether to accept the targets, and choose whether to enter the agreement. We model emissions decisions using a neutrally framed public bads game (except for the term "disaster"), where the probability of disasters increases exponentially with production and disaster risk is idiosyncratic.

In the experiment, participants were matched in groups of three and simultaneously chose their production $q_i \in [0,40]$, for $i \in \{1, 2, 3\}$. Total production by all three participants increased the risk of a climate disaster (r_i), calculated as $r_i(q_1, q_2, q_3) = \min\left\{\left(\frac{q_1+q_2+q_3}{100}\right)^2, 1\right\}$. In other words, if the total production exceeded 100 units, the disaster was guaranteed to occur; otherwise, the probability increased in the total production. Although all participants shared the same probability of facing a climate disaster, the occurrence of a disaster was idiosyncratic: a separate disaster draw was made for each participant. This design choice allows us to capture the fact that although climate change is a common global threat, climate disasters do not strike all the regions of the Earth at the same time. Participants i would earn $\pi_i = 12q_i$ if a disaster did not occur and nothing if a disaster occurred. Expected earnings are therefore calculated as $E[\pi_i] = 12q_i\left(1 - \min\left\{\left(\frac{q_1+q_2+q_3}{100}\right)^2, 1\right\}\right)$. The game features a similar tradeoff between individual and collective interest as the other social dilemma games, but it also includes additional elements that capture the institutional context and thus increase the external validity of our results. For example, the game was framed in terms of the creation of public bads, rather than public goods, the costs of emissions were convex, and emissions increased the probability of a disaster instead of lowering the payoffs for sure.

Table 1. Experimental design and treatments.

	Baseline	**+ Pledge-and-review**	**+ Penalties**
No evaluation	*Baseline*	*Pledge*	*Pledge-Penalty*
Peer evaluation	*Baseline-Evaluation*	*Pledge-Evaluation*	*Pledge-Penalty-Evaluation*

We used a 2x3 factorial design, as shown in Table 1. In each treatment, the game was played for 18 rounds. The first three rounds were identical in all six treatments; we label these rounds as "business-as-usual", as they measure production in the absence of additional mechanisms. In the *Baseline* treatment, the remaining 15 rounds were also the same as the first three. In the other five treatments, further instructions were provided at the start of the 15 rounds, explaining the additional mechanism that simulates climate agreements (see the instructions in Appendix S3). The first added element is the pledge-and-review process, either without financial penalties ("*Pledge*" treatments) or with penalties for those who produce more than pledged ("*Pledge-Penalty*" treatments). The second element is peer evaluation: in the "*Peer evaluation*" treatments, participants evaluated their peers; in the other three treatments, there was no peer evaluation.

In the *Pledge* treatments, the 15 rounds were divided into five blocks of 3 rounds. At the start of each block, participants simultaneously selected their own targets, pledging to produce no more than the target in each of the 3 rounds in that block. Everyone in the group observed each other's pledges and voted on whether to confirm or reject the agreement. If the majority rejected the agreement, the three players simultaneously revised the pledges and subsequently voted again. After three rejections, no pledges were in effect and the game continued as in the *Baseline* treatment. If the majority voted to accept the pledges, an agreement was formed and participants could individually choose whether to join it. Participants who joined the agreement were reminded about their pledge at the start of each round and everyone in the group saw whether their output exceeded the pledge at the end of each round. There were no penalties for exceeding the pledged amount. The pledge-and-review procedure with endogenous pledges followed by a vote and a ratification stage is consistent with the design of the Paris Agreement and recent theoretical work (Harstad, 2023).

The *Pledge-Penalty* treatments were identical to the *Pledge* treatments, but participants who violated the pledge paid a penalty of 8 experimental currency units for each unit above the pledge. The collected revenue was evenly redistributed to all the participants who joined the agreement. Participants who did not join the agreement did not pay any penalties and received no payouts from the collected revenue. The penalty rate was set to enforce the socially optimal amount if all participants pledged to produce below the socially optimal level and joined the agreement (see Proposition 1 in Appendix S2).

In the *Peer evaluation* treatments, we introduced a contest for environmental reputation. At the end of each round, participants picked their favorite group member and assigned them 1, 2, or 3 reputation points. The assignment of points was zero-sum: for each point assigned to their favorite group member, the other member lost the same number of points. Following previous work on rankings and alliances (DeScioli & Kimbrough, 2019), the reputation contest has a zero-sum feature so that only one country can be at the top. Participants could also choose to not reassign any points. At the end of each round, participants saw the reputation ranking, reputation points, change in reputation ranking (indicated by an up-or-down arrow), and change in reputation points for each group member**.** After the game, the group member with the most reputation points had a tree planted in their name by the organization One Tree Planted in Indonesia.

Experimental Procedures

We ran the experiment with 371 participants, recruited from the subject pool at the National University of Singapore (NUS) using ORSEE (Greiner, 2015). The experiment was approved by the

Institutional Review Board at the National University of Singapore (reference number: S-18-363) and run online using the LIONESS Lab platform (Giamattei et al., 2020). The experiments were conducted online because the physical laboratory was closed due to the COVID-19 restrictions in Singapore when we ran the study. 60 participants took part in *Baseline* and *Pledge-Penalty* treatments; 68 in *Pledge*; 64 in the *Baseline-Evaluation*; 63 in *Pledge-Evaluation* and *Pledge-Penalty-Evaluation*. Some participants disconnected during the experiment; when that occurred, the experiment was terminated for the entire group and participants received their accumulated earnings, as it was not possible to proceed with only two players. In total, 51 participants completed the entire experiment in each treatment with peer evaluation and 54 completed it in each treatment without peer evaluation. Our analysis includes the data up to the point of termination, but the results do not change if we exclude the groups that terminated early (doing so removes 4.6% of the observations). Table A8 replicates the main results shown in Table 2, excluding the incomplete data[2].

Participants had to correctly answer four quiz questions to start the experiment (see Appendix S3). Participants received the sum of earnings from all 18 rounds, which amounted to an average of 10.70 Singapore dollars (at the time of the experiment, the exchange rate was 1 SGD = 0.75 USD). The earnings were transferred to the bank accounts of the participants. The experiment on average took 30 minutes.

The relationship between the total production and disaster risk was explained using a table and a figure (see Appendix S3). We also provided two examples, which showed the disaster risk and potential payoffs for a hypothetical set of production decisions. The instructions were neutrally framed (see Appendix S3), avoiding terms related to emissions or climate change, although we referred to the outcome of losing income for the current round as a “disaster”, to facilitate

[2] Both methods of dealing with attrition have limitations. Keeping the data up to the point of termination means that the treatments will have a different number of observations for each round. This might distort the average production, as the production tends to increase over time (Figure 1). Removing all the data from the groups that disconnected solves this issue but could bias the results if attrition is endogenously determined by the history of play.

comprehension. In the treatments with penalties for exceeding the pledged production, we used the term "tax" as a neutral term that is commonly employed in the real world in the context of public bads (for example, the "carbon tax" in the European Union). At the end of each round, participants received detailed information about the disaster risk, potential earnings, and realized earnings of each group member (see screenshots in Appendix S4). In treatments with pledge-and-review, participants were also informed about who joined the agreement, what pledges they made, and whether production exceeded the pledge, and received detailed information about how the earnings of each participant were calculated, including any penalties paid.

In every round, participants had 60 seconds to make each decision (choose production, pledges, vote for pledges, join the agreement) and 60 seconds to view the feedback screen. If no decision was made during this time, we imposed a conservative default action (producing nothing, pledging the highest possible amount, rejecting the pledge, or not joining the agreement). Excluding the participants who dropped out, the production decision was not made only five times out of 5,669 decisions (less than 0.1%). The actions imposed by default were excluded from the analysis. On average, it took participants 7.2 seconds to make the production decision, ranging from 6.1 seconds in the *Baseline* to 8.2 seconds in *Pledge-Penalty-Evaluation*. Less than 1% of the decisions took more than 50 seconds to make.

Game-Theoretic Predictions

We formulate hypotheses about output in each treatment by computing the symmetric subgame-perfect Nash equilibrium (see Appendix S2 for details). In short, we show that the financial penalties are predicted to have no effect on output because the agreements are voluntary, thus participants who anticipate having to pay a penalty would not join an agreement or would join it with unambitious pledges. Therefore, theory predicts that climate agreements will not reduce emissions, as the free-riding problem is merely shifted from the production decision to the decision about joining the agreement.

In the *Baseline* treatments, there is a unique stage game Nash equilibrium in which each participant produces 25.8 units and the disaster risk is 60%. It is socially optimal for each participant to produce 19.2 units, which would lower the disaster risk to 33%. In the *Pledge, Base-Evaluation,* and *Pledge-Evaluation* treatments, pledges and peer evaluation are cheap talk, therefore the predictions are identical to the *Baseline*.

In the *Pledge-Penalty* and *Pledge-Penalty-Evaluation* treatments, there are no symmetric subgame-perfect Nash equilibria in which total production in the *Pledge-Penalty* treatments is lower than in the *Baseline* (see Proposition 2 in Appendix S2). Financial penalties fail to reduce emissions because participants would be individually better off not joining the agreements if they anticipate having to pay the penalties, even though such agreements, if joined by everyone, would eliminate the tradeoff between maximizing individual and collective payoffs. We calibrated the penalty rate so that if players decided to join the agreement and make sufficiently ambitious pledges, they would produce the socially efficient amount (see Proposition 1 in Appendix S2). However, because joining the agreement and making ambitious pledges was voluntary, players would either not join or join with unambitious pledges and consequently would not reduce production compared to the *Baseline* treatment.

Behavioral Predictions

Game theory predicts that none of the manipulations will decrease production below the *Baseline* treatment. However, treatment effects might appear if participants care about behavioral factors, in addition to monetary earnings.

The pledge-and-review procedure could reduce subsequent production for two reasons. First, it has been shown that non-binding commitments can facilitate sustained pro-environmental behaviors (Lokhorst et al., 2013). The theory of commitment proposes several methods to make commitments more effective, such as making them explicit, public, and freely chosen (Joule et al., 2007), as they are in our experiment. Second, the pledge-and-review procedure could reduce production by providing assurance for conditional cooperators. It is known that many participants in social dilemma games are willing to cooperate if they know that other group members will do so as well (Fischbacher et al., 2001). Pledges and votes over the submitted pledges could assure the conditional cooperators, who would use the pledge-and-review mechanism as a coordination device to achieve the socially optimal outcome (Dannenberg, 2016). We therefore hypothesize that production will be lower in the *Pledge* treatment compared to the *Baseline* treatment.

Behavioral predictions about the effect of financial penalties are ambiguous. On one hand, penalties facilitate enforcement, assuring conditional cooperators that the other group members will stick to their pledges. As a result, penalties create a mechanism that provides a stronger commitment device. On the other hand, penalties might lead to weak agreements by discouraging participation and repressing ambitious pledges. In addition, monetary penalties might crowd out the intrinsic motivation to reduce production (Gneezy & Rustichini, 2000), shifting choices towards the subgame-perfect Nash equilibrium predictions.

We also consider the behavioral effect of non-monetary peer evaluation. If participants care about their reputation and higher emissions reduce reputation, the peer evaluation mechanism should lower emissions. Previous work has found some evidence that people are willing to trade off money for status. For instance, activating people's motivation for status induces them to choose environmentally friendly products over luxury products (Griskevicius et al., 2010). Moreover, theoretical work in economics (Loch et al., 2000) and evolutionary psychology (Henrich et al., 2015) showed that competing for status can induce people to cooperate. If rank in a group depends on environmentally friendly behavior, to the extent that group members seek social status over money, highly-ranked members may attract support (DeScioli & Kurzban, 2013) and invite imitation (Eckel et al., 2010; Koessler, 2019), fostering a virtuous cycle of ever-lower emissions.

Specific predictions about the effect of peer evaluation depend on the non-monetary costs of receiving a low evaluation, as well as on the criteria used to evaluate peers. If evaluations are based solely on emissions, regardless of the pledge and the gap between the pledge and emissions, peer evaluation should decrease emissions by the same amount in all three treatments with peer evaluation. However, evaluation could also be based on pledges, as those who exceed their pledge might receive a lower score than those who never made such a promise. Evaluation might also interact with financial penalties, as participants might not wish to assign low evaluations to those who were already financially punished for exceeding their pledge; in other words, monetary penalties might crowd out non-monetary punishment (Gneezy & Rustichini, 2000).

Results

Production and disaster risk

Table 2 summarizes the main variables in all six treatments from rounds 4-18 (the first three rounds that measure business-as-usual production are excluded). We find that average production in all treatments exceeded the social optimum (19.2) but was below the equilibrium prediction (25.8). Introducing climate pledges (with or without penalties) and peer evaluation did not affect the average production or the disaster risk (Mann-Whitney U test, two-sided $p > 0.46$ for all pairwise comparisons). To model the panel structure of the data, we analyzed the results using GLS regressions with a random effect at the participant level and standard errors clustered at the group level (Appendix S1, Table A1). We again found no significant difference between the *Baseline* and the other five treatments, either in terms of production (models 1-3) or disaster risk (models 4-6). The results did not change when we additionally controlled for the average production in the first 3 rounds (models 2 and 5) or age and gender (models 3 and 6). Finally, looking at decisions over time, Figure 1 shows that production is increasing and there is no difference in trends between the treatments.

Table 2. Average and standard deviation (in parentheses) of the main variables, by treatment.

	Production	Disaster Risk	Pledge	Participation	Ambition	Compliance
Baseline	23.5 (8.3)	0.51 (0.19)				

Baseline-Evaluation	23.2 (6.6)	0.50 (0.18)				
Pledge	23.6 (6.6)	0.52 (0.17)	25.4 (6.7)	0.85 (0.36)	33.1%	0.90 (0.29)
Pledge-Evaluation	23.5 (6.8)	0.51 (0.19)	25.7 (6.1)	0.89 (0.31)	16.3%	0.94 (0.24)
Pledge-Penalty	24.0 (5.8)	0.53 (0.15)	26.3 (6.2)	0.80 (0.40)	18.7%	0.98 (0.15)
Pledge-Penalty-Evaluation	23.6 (7.2)	0.53 (0.23)	26.0 (7.9)	0.67 (0.47)	13.1%	0.96 (0.20)

Note. Data from rounds 4-18. Participation is calculated as the share of participants who joined the agreement. Ambition is calculated as the percentage of pledges that were below the business-as-usual production (participant's average production in the first 3 rounds of the game). Compliance is calculated as the fraction of production choices that are below the pledged production. Compliance and ambition are conditional on joining the agreement.

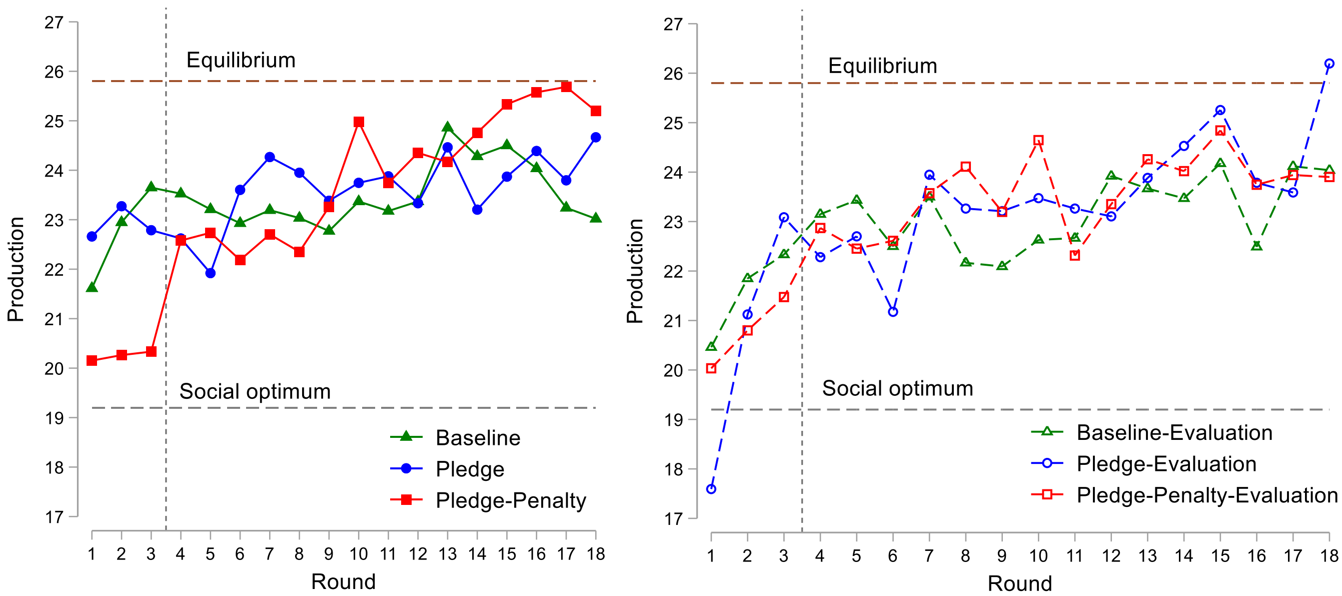

Figure 1. Mean production across rounds in treatments without peer evaluation (left panel) and with peer evaluation (right panel), compared to the equilibrium prediction and the social optimum. The vertical dashed line shows when the treatment's mechanism was introduced.

Participation, ambition, and compliance

Next, we study why the pledge-and-review mechanism failed to reduce production. Successful climate agreements require sufficient participation, ambitious pledges, and compliance. Therefore, we compare the treatments by (1) whether participants joined agreements, (2) whether those who joined made ambitious pledges, and (3) whether joiners complied with their pledges. We measure participation by how frequently participants join agreements. We measure ambition by the average pledge (lower pledges are more ambitious) and by measuring how many of those who join the agreement pledge to produce less than their business-as-usual level (average production in the first three rounds). Finally, we measure compliance by calculating how frequently participants produce more than they pledged and by how close their production is to the pledged amount.

Participation. In the *Pledge* treatment, 85% of participants joined an agreement. Adding penalties for exceeding the pledge decreased participation, although the decrease is significant only when penalties were combined with peer evaluation (participation decreased from 89% in *Pledge-Evaluation* to 67% in *Pledge-Penalty-Evaluation*; MWU $p = 0.03$ when not accounting for multiple hypothesis testing (MHT), $p = 0.13$ when correcting for MHT[3]). Peer evaluation did not affect joining rates (MWU $p = 0.324$ for the difference between *Pledge-Penalty* and *Pledge-Penalty-Evaluation*; $p = 0.42$ between *Pledge* and *Pledge-Evaluation*). We also examine the effect of participation on production. Since joining an agreement was endogenous, the correlation between production and the decision to join could be driven by self-selection. To control for possible self-selection, we regress production on joining, controlling for business-as-usual production (average production in the first three rounds). Pooling all the treatments together, we find that participants who join the agreement produce significantly less than those who do not, even when controlling for their business-as-usual production (Appendix S1, Table A2). This reduction is primarily driven by the *Pledge* treatment. In the other treatments, the reduction was directionally similar but not statistically significant.

Ambition. We assess the ambition of pledges in two ways. First, we measure ambition by comparing the average implemented pledge (defined as the pledge that was accepted by the majority

[3] In this section, we apply the conservative Holm correction to the Mann-Whitney *U* test, taking to account that there are 4 possible pairwise comparisons in the 2x2 design that is used for the treatments with a pledge-and-review mechanism. The Bonferroni correction yields identical results, except for the case explained in the next footnote.

and the participant who made the pledge joined the agreement), averaged across all the rounds and all participants in the group. We find no significant difference in average pledges between the four pledge-and-review treatments (all pairwise MWU $p > 0.3$). However, average pledges fail to account for outliers and business-as-usual production. For instance, a pledge to produce 25 units by a participant who was initially producing 30 units could be regarded as ambitious, whereas the same pledge made by a participant who was producing 20 units could be considered unambitious. Therefore, we use an alternative measure of ambition: the fraction of participants who pledged to produce less than their business-as-usual production (average production in the first three rounds of the game). Using this measure, we find that most of the pledges are unambitious (see Table 2): the treatment with the highest level of ambition is *Pledge*, where 33.1% of the pledges are ambitious. This figure drops to 13.1% in the *Pledge-Penalty-Evaluation* treatment, whereas treatments with only financial penalties or only peer evaluation fall in between. We also find that introducing peer evaluation reduces ambition, but only when there are no financial penalties (MWU $p = 0.0491$ for the difference between *Pledge* and *Pledge-Evaluation;* $p = 0.1964$ when corrected for MHT).

Compliance. Most of the participants complied with their pledges, perhaps because pledges were often not ambitious. Compliance ranged from 90.9% in the *Pledge* treatment to 97.5% in the *Pledge-Penalty* treatment (Table 2). Financial penalties improved compliance, both with peer evaluation (MWU $p = 0.0237$; $p = 0.0711$ when corrected for MHT[4]) and without it (MWU $p = 0.0113$; $p = 0.0452$ when corrected for MHT). We examine compliance further by regressing production on the chosen pledge. Since the correlation between these variables could be driven by lower pledges chosen by the participants who planned to produce less, we additionally control for the production in the three business-as-usual rounds. We find that in all treatments, participants with lower pledges also produce less, even when controlling for their business-as-usual production. We

[4] If we use the Bonferroni instead of the Holm correction, $p = 0.0948$.

also find that the effect of pledges on production is stronger when there are penalties for exceeding the pledge (Appendix S1, Table A4).

We find very similar results using random-effects GLS and probit models (Appendix S1, Table A3), instead of the non-parametric tests. Penalties and peer evaluation do not affect the willingness to join the agreement (model 1 in Table A3). They also do not affect average pledges (model 2). However, peer evaluation reduces the frequency of ambitious pledges, measured by being below business-as-usual production (model 3). We also find that penalties for exceeding a pledge increase compliance, measured by the frequency of production below the pledge (model 4).

Overall, we conclude that in all treatments, most participants join agreements and comply with their pledges. However, most pledges are not ambitious and thus fail to reduce production below the business-as-usual level or below the *Baseline* treatment. Financial penalties improve compliance rates and peer evaluation lowers ambition, but neither affects production.

Peer evaluation

In treatments with peer evaluation, the logic of 'naming and shaming' invited by the Paris Agreement suggests that participants would rank lower those peers who created the negative externality by producing more. Alternatively, evaluation could be based not on the absolute production level, but on the change in production compared to the previous round. We study the peer evaluation decisions by regressing the number of reputation points that a participant received on their previous production and its change. We find that those who produce more receive significantly fewer reputation points from their peers than those who produce less (Appendix S1, Table A5, models 1-5). We also find that the number of received points does not depend on the change in production compared to the previous round (Appendix S1, Table A5, models 1-3). We also test whether more ambitious pledges improve peer evaluations; they do not (models 4-5 in Table A5). In the treatment without penalties, unambitious pledges even boosted the number of received points (model 4 in Table A5).

Next, we study how participants react to peer evaluation. We regress the change in production from the previous round on the number of reputation points that the participant received in previous rounds, controlling for their existing accumulated reputation and business-as-usual production in the first three rounds (Appendix S1, Table A6). We must also account for regression to the mean, as lower evaluations are given to those who produce more (see previous paragraph), and such participants are more likely to subsequently lower their production. We find that when the regression to the mean is accounted for by controlling for the production in the previous round, the received reputation points do not affect subsequent production (models 4-6 in Table A6). If previous production was not included, the results would show that the lower-ranked participants subsequently decreased production in all treatments with peer evaluation (models 1-3 in Table A6).

Reaction to disasters

Finally, we test whether participants who experience a climate disaster subsequently lower their production. To account for reverse causality, we regress production on the incidence of climate disaster in the previous round, controlling for the previous production and disaster risk (Table A7 in Appendix S1). This approach allows us to capture the pure effect of the incidence of disasters. Overall, we find that participants who experience a disaster subsequently reduce their production by on average 0.4 units, an effect that was primarily driven by the *Pledge* and *Pledge-Penalty* treatments.

Discussion

We designed an experiment to test whether climate agreements based on pledges, financial penalties, and peer evaluation can successfully curb emissions. We found that they could not: even though agreements enjoyed broad membership and pledges were often met, the submitted pledges were usually not ambitious. Consequently, even the participants who met their pledges did not reduce the emissions below the business-as-usual level. Both monetary and non-monetary penalties improved compliance, but it was already quite high (above 90%) even without penalties.

Game-theoretic analysis shows that if participants were required to join the agreements and chose ambitious pledges, emissions in the treatment with financial penalties would drop to the socially efficient level. However, when given the chance to opt out of such agreements, participants have the incentive to do so and produce as much as in the treatment with no financial penalties. The experimental results support these predictions and highlight the difficulty of climate accords based on decentralized and voluntary agreements: when commitments and participation are voluntary, parties can avoid financial and reputational repercussions by not joining or by making unambitious pledges. Our experiment thus adds to the growing literature on institutional choice in social dilemmas (Dannenberg & Gallier, 2020) pointing to the shortcomings of both "broad but shallow" and "narrow but deep" climate agreements (Aldy et al., 2003). Namely, "broad but shallow" agreements are vulnerable to the possibility that joiners will mostly continue with business as usual. In contrast, "narrow but deep" agreements face the risk of failing to generate a sufficient drop in total emissions. For both types of agreements, failure to achieve group targets is a common shortcoming.

Despite the overall failure of pledge-and-review institutions in our experiment, we can draw several lessons for the design of agreements. One important takeaway is that neither monetary nor non-monetary penalties improve the ambition of pledges, and thus fail to reduce production. Monetary penalties failed because they could be avoided by making unambitious pledges or refusing to join the agreement. Non-monetary penalties due to reputation loss cannot be avoided, but we found that the participants who lost reputation points did not reduce subsequent production. A global contest for environmental reputation called for by the Paris Agreement might therefore have little effect on emissions. A second lesson is that predicating pledges on peer evaluation can be a double-edged sword. If failure to meet pledges is an important concern, a greater role for peer evaluation may invite more conservative pledges instead of more ambitious ones. This result resonates with recent research about Americans' preferences for climate pledges, which found greater public support for pledges that do not over- or under-promise (Tingley & Tomz, 2020).

We make several contributions to the previous literature on the design of climate treaties. We are the first to experimentally study whether monetary penalties for exceeding the pledged amount can reduce emissions, as envisaged by the Kyoto Protocol. We are also the first to study the interaction between monetary and non-monetary penalties. In terms of methodology, we created a framework for studying decentralized and voluntary climate change agreements that could be extended to evaluate other elements of international agreements in future research. We model the probabilistic nature of climate disasters in a novel game that more accurately describes the consequences of emissions than the commonly used public goods game. This tractable game allows us to study, both theoretically and experimentally, how the compliance mechanism affects production.

The present research comes with some limitations. Real agreements are more complex and involve more parties compared to the stylized three-player model that we used. As a result, our findings might not extend to the climate change negotiations that involve almost two hundred nations, and future research should investigate how the effectiveness of the climate treaties changes as the number of parties increases. Also, real agreements are the result of lengthy face-to-face negotiations, whereas participants in our study played anonymously behind computers. Thus, the reputation built during the game was confined to their avatar and did not affect their reputation outside of the lab. In contrast, when global actors vie for climate reputation, their country's image is under much more intense scrutiny for a longer time and with much larger consequences. Real agreements are also more complicated because countries differ in production capacity and historical emissions (Del Ponte et al., 2023); future research could extend our findings to an asymmetric setting. Also, the penalty rates were set exogenously; future research could model negotiations over the penalty rates to test the performance of mechanisms with endogenously determined penalty rates. Future research could also evaluate the mechanisms with a different baseline game, such as threshold or linear public goods games rather than the public bads game that we used. Previous literature found that the effectiveness of interventions differs between the public bads and the public goods games (Abatayo & Li, 2024) and it would be useful to know if our results replicate in different settings.

Future research could also evaluate how group preferences are aggregated to determine the parameters of the climate agreements. Currently, the rules of procedure at the United Nations Framework Convention on Climate Change are not formally defined, but decisions generally require a loosely understood general agreement (or consensus) rather than strict unanimity (Rietig et al., 2023). This procedure has been criticized for being inefficient and ineffective compared to majority voting (Kemp, 2016; Vihma, 2015). To create favorable conditions for the climate agreements, our study used majority voting, finding that it was not sufficient for the agreements to be effective. It would be interesting to explore other procedural rules including unanimity, perhaps additionally varying the framing of the procedure (either requiring everyone to vote in favor or proceeding with the agreement unless someone explicitly objects).

Taken together, our experiment points to the perils of focusing on formal compliance with international agreements instead of concentrating efforts on reaching climate targets. Since the compliance mechanisms could not decrease emissions in the laboratory, it seems unlikely that they would work better in practice, where financial penalties are difficult to enforce (Hovi et al., 2007; Victor, 2011), and the climate agenda has rarely been a concrete priority for global leaders (Christoff, 2010; Conrad, 2012), even amid rising public outcry about the urgency of addressing climate change (Evensen, 2019).

References

Abatayo, A. L., & Li, T. (2024). Public goods and bads with vulnerable individuals: How information and social nudges change behaviour. *Canadian Journal of Economics/Revue Canadienne d'économique*, *57*(2), 556–587.

Aldy, J. E., Barrett, S., & Stavins, R. N. (2003). Thirteen plus one: A comparison of global climate policy architectures. *Climate Policy*, *3*(4), 373–397.

Alt, M., Gallier, C., Kesternich, M., & Sturm, B. (2023). Collective minimum contributions to counteract the ratchet effect in the voluntary provision of public goods. *Journal of Environmental Economics and Management*, *122*, 102895.

Balliet, D. (2010). Communication and cooperation in social dilemmas: A meta-analytic review. *Journal of Conflict Resolution*, *54*(1), 39–57.

Barrett, S. (2003). *Environment and statecraft: The strategy of environmental treaty-making: The strategy of environmental treaty-making*. OUP Oxford.

Barrett, S., & Dannenberg, A. (2016). An experimental investigation into 'pledge and review' in climate negotiations. *Climatic Change*, *138*(1–2), 339–351. https://doi.org/10.1007/s10584-016-1711-4

Bochet, O., & Putterman, L. (2009). Not just babble: Opening the black box of communication in a voluntary contribution experiment. *European Economic Review*, *53*(3), 309–326.

Chaudhuri, A. (2011). Sustaining cooperation in laboratory public goods experiments: A selective survey of the literature. *Experimental Economics*, *14*, 47–83.

Chávez, C. A., Murphy, J. J., Quezada, F. J., & Stranlund, J. K. (2023). The endogenous formation of common pool resource coalitions. *Journal of Economic Behavior & Organization*, *211*, 82–102.

Cherry, T. L., Kallbekken, S., Sælen, H., & Aakre, S. (2021). Can the Paris Agreement deliver ambitious climate cooperation? An experimental investigation of the effectiveness of pledge-and-review and targeting short-lived climate pollutants. *Environmental Science & Policy*, *123*, 35–43.

Cherry, T. L., & McEvoy, D. M. (2013). Enforcing compliance with environmental agreements in the absence of strong institutions: An experimental analysis. *Environmental and Resource Economics*, *54*, 63–77.

Christoff, P. (2010). Cold climate in Copenhagen: China and the United States at COP15. *Environmental Politics*, *19*(4), 637–656.

Conrad, B. (2012). China in Copenhagen: Reconciling the "Beijing climate revolution" and the "Copenhagen climate obstinacy." *The China Quarterly*, *210*, 435–455.

Dannenberg, A. (2016). Non-binding agreements in public goods experiments. *Oxford Economics Papers*, *68*(1), 279–300. https://doi.org/10.1093/oep/gpv048

Dannenberg, A., & Gallier, C. (2020). The choice of institutions to solve cooperation problems: A survey of experimental research. *Experimental Economics*, *23*(3), 716–749.

Del Ponte, A., Masiliunas, A., & Lim, N. (2023). Information about historical emissions drives the division of climate change mitigation costs. *Nature Communications*.

DeScioli, P., & Kimbrough, E. O. (2019). Alliance formation in a side-taking experiment. *Journal of Experimental Political Science*, *6*(1), 53–70.

DeScioli, P., & Kurzban, R. (2013). A solution to the mysteries of morality. *Psychological Bulletin*, *139*(2), 477.

du Pont, Y. R., & Meinshausen, M. (2018). Warming assessment of the bottom-up Paris Agreement emissions pledges. *Nature Communications*, *9*(1), 1–10.

Dugar, S. (2013). Non-monetary incentives and opportunistic behavior: Evidence from a laboratory public good game. *Economic Inquiry*, *51*(2), 1374–1388.

Eckel, C. C., Fatas, E., & Wilson, R. (2010). Cooperation and status in organizations. *Journal of Public Economic Theory*, *12*(4), 737–762.

Evensen, D. (2019). The rhetorical limitations of the# FridaysForFuture movement. *Nature Climate Change*, *9*(6), 428–430.

Faillo, M., Grieco, D., & Zarri, L. (2020). The impact of peer ratings on cooperation: The role of information and cost of rating. *Journal of Public Economic Theory*, *22*(2), 408–432.

Falkner, R. (2016). The Paris Agreement and the new logic of international climate politics. *International Affairs*, *92*(5), 1107–1125.

Fehr, E., & Gächter, S. (2000). Fairness and retaliation: The economics of reciprocity. Journal of Economic Perspectives. *The Journal of Economic Perspectives*, *14*(3), 159–181. https://doi.org/10.1257/jep.14.3.159

Feige, C., Ehrhart, K.-M., & Krämer, J. (2018). Climate Negotiations in the Lab: A Threshold Public Goods Game with Heterogeneous Contributions Costs and Non-binding Voting. *Environmental and Resource Economics*, *70*, 343–362. https://doi.org/10.1007/s10640-017-0123-x

Fischbacher, U., Gächter, S., & Fehr, E. (2001). Are people conditionally cooperative? Evidence from a public goods experiment. *Economics Letters*, *71*(3), 397–404.

Gallier, C. (2020). Democracy and compliance in public goods games. *European Economic Review*, *121*, 103346.

Gallier, C., & Sturm, B. (2021). The ratchet effect in social dilemmas. *Journal of Economic Behavior & Organization*, *186*, 251–268.

Giamattei, M., Yahosseini, K. S., Gächter, S., & Molleman, L. (2020). LIONESS Lab: A free web-based platform for conducting interactive experiments online. *Journal of the Economic Science Association*, *6*(1), 95–111.

Gneezy, U., & Rustichini, A. (2000). Pay enough or don't pay at all. *The Quarterly Journal of Economics*, *115*(3), 791–810.

Greiner, B. (2015). Subject pool recruitment procedures: Organizing experiments with ORSEE. *Journal of the Economic Science Association*, *1*(1), 114–125.

Griskevicius, V., Tybur, J. M., & Van den Bergh, B. (2010). Going green to be seen: Status, reputation, and conspicuous conservation. *Journal of Personality and Social Psychology*, *98*(3), 392.

Halvorssen, A., & Hovi, J. (2006). The nature, origin and impact of legally binding consequences: The case of the climate regime. *International Environmental Agreements: Politics, Law and Economics*, *6*(2), 157–171.

Harstad, B. (2023). Pledge-and-review bargaining. *Journal of Economic Theory*, *207*, 105574.

Held, D., & Roger, C. (2018). Three Models of Global Climate Governance: From Kyoto to Paris and Beyond. *Global Policy*, *9*(4), 527–537. https://doi.org/10.1111/1758-5899.12617

Henrich, J., Chudek, M., & Boyd, R. (2015). The Big Man Mechanism: How prestige fosters cooperation and creates prosocial leaders. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *370*(1683), 20150013.

Hovi, J., Froyn, C. B., & Bang, G. (2007). Enforcing the Kyoto Protocol: Can punitive consequences restore compliance? *Review of International Studies*, *33*(3), 435–449.

IPCC. (2023). Climate Change 2023: Synthesis Report. A Report of the Intergovernmental Panel on Climate Change. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the

Intergovernmental Panel on Climate Change. *The Intergovernmental Panel on Climate Change (IPCC): Geneva, Switzerland*.

Jacquet, J., & Jamieson, D. (2016). Soft but significant power in the Paris Agreement. *Nature Climate Change*, *6*(7), 643–646.

Joule, R., Girandola, F., & Bernard, F. (2007). How can people be induced to willingly change their behavior? The path from persuasive communication to binding communication. *Social and Personality Psychology Compass*, *1*(1), 493–505.

Kemp, L. (2016). Framework for the future? Exploring the possibility of majority voting in the climate negotiations. *International Environmental Agreements: Politics, Law and Economics*, *16*, 757–779.

Koessler, A.-K. (2019). *Setting new behavioural standards: Sustainabilty pledges and how conformity impacts their outreach*. ZBW – Leibniz Information Centre for Economics, Kiel, Hamburg. http://hdl.handle.net/10419/195048

Kosfeld, M., Okada, A., & Riedl, A. (2009). Institution formation in public goods games. *American Economic Review*, *99*(4), 1335–1355.

Ledyard, J. O. (1995). Public Goods: A Survey of Experimental Research. In *Handbook of Experimental Economics* (Kagel, John, and Roth Alvin). Princeton University Press.

Lippert, S., & Tremewan, J. (2021). Pledge-and-review in the laboratory. *Games and Economic Behavior*.

Loch, C. H., Huberman, B. A., & Stout, S. (2000). Status competition and performance in work groups. *Journal of Economic Behavior & Organization*, *43*(1), 35–55.

Lokhorst, A. M., Werner, C., Staats, H., van Dijk, E., & Gale, J. L. (2013). Commitment and behavior change: A meta-analysis and critical review of commitment-making strategies in environmental research. *Environment and Behavior*, *45*(1), 3–34.

Masclet, D., Noussair, C., Tucker, S., & Villeval, M.-C. (2003). Monetary and nonmonetary punishment in the voluntary contributions mechanism. *American Economic Review*, *93*(1), 366–380.

McEvoy, D. M., Haller, T., & Blanco, E. (2022). The Role of Non-Binding Pledges in Social Dilemmas with Mitigation and Adaptation. *Environmental and Resource Economics*, *81*(4), 685–710.

Nentjes, A., & Klaassen, G. (2004). On the quality of compliance mechanisms in the Kyoto Protocol. *Energy Policy*, *32*(4), 531–544.

Peeters, R., & Vorsatz, M. (2013). Immaterial rewards and sanctions in a voluntary contribution experiment. *Economic Inquiry*, *51*(2), 1442–1456.

Rietig, K., Peringer, C., Theys, S., & Censoro, J. (2023). Unanimity or standing aside? Reinterpreting consensus in United Nations Framework Convention on Climate Change negotiations. *International Environmental Agreements: Politics, Law and Economics*, *23*(3), 221–234.

Roelfsema, M., van Soest, H. L., Harmsen, M., van Vuuren, D. P., Bertram, C., den Elzen, M., Höhne, N., Iacobuta, G., Krey, V., & Kriegler, E. (2020). Taking stock of national climate policies to evaluate implementation of the Paris Agreement. *Nature Communications*, *11*(1), 1–12.

Schmidt, K. M., & Ockenfels, A. (2021). Focusing climate negotiations on a uniform common commitment can promote cooperation. *Proceedings of the National Academy of Sciences*, *118*(11).

Tavoni, A., Dannenberg, A., Kallis, G., & Löschel, A. (2011). Inequality, communication, and the avoidance of disastrous climate change in a public goods game. *Proceedings of the National Academy of Sciences*, *108*(29), 11825–11829.

Tingley, D., & Tomz, M. (2020). International commitments and domestic opinion: The effect of the Paris Agreement on public support for policies to address climate change. *Environmental Politics*, *29*(7), 1135–1156.

Tyran, J., & Feld, L. P. (2006). Achieving compliance when legal sanctions are non-deterrent. *Scandinavian Journal of Economics*, *108*(1), 135–156.

Victor, D. G. (2011). *The collapse of the Kyoto Protocol and the struggle to slow global warming*. Princeton University Press.

Vihma, A. (2015). Climate of consensus: Managing decision making in the UN climate change negotiations. *Review of European, Comparative & International Environmental Law*, *24*(1), 58–68.

Vollan, B., Landmann, A., Zhou, Y., Hu, B., & Herrmann-Pillath, C. (2017). Cooperation and authoritarian values: An experimental study in China. *European Economic Review*, *93*, 90–105.

Young, O. R. (2016). The Paris Agreement: Destined to Succeed or Doomed to Fail? *Politics and Governance*, *4*(3), 124–132. https://doi.org/10.17645/pag.v4i3.635