Barrett, Scott; Dannenberg, Astrid

Working Paper

Negotiating to Avoid Gradual versus Dangerous Climate Change: An Experimental Test of Two Prisoners' Dilemma

CESifo Working Paper, No. 4573

Provided in Cooperation with:
Ifo Institute – Leibniz Institute for Economic Research at the University of Munich

Suggested Citation: Barrett, Scott; Dannenberg, Astrid (2014) : Negotiating to Avoid Gradual versus Dangerous Climate Change: An Experimental Test of Two Prisoners' Dilemma, CESifo Working Paper, No. 4573, Center for Economic Studies and ifo Institute (CESifo), Munich

This Version is available at:
https://hdl.handle.net/10419/89722

The international platform of Ludwig-Maximilians University's Center for Economic Studies and the Ifo Institute

CESifo Working Papers
www.cesifo.org/wp

Negotiating to Avoid "Gradual" versus "Dangerous" Climate Change: An Experimental Test of Two Prisoners' Dilemmas

Scott Barrett
Astrid Dannenberg

CESIFO WORKING PAPER NO. 4573
CATEGORY 10: ENERGY AND CLIMATE ECONOMICS
JANUARY 2014

CESifo
Center for Economic Studies & Ifo Institute

CESifo *Working Paper No. 4573*

Negotiating to Avoid "Gradual" versus "Dangerous" Climate Change: An Experimental Test of Two Prisoners' Dilemmas

Abstract

According to the Framework Convention on Climate Change, global collective action is needed to stabilize "greenhouse gas concentrations in the atmosphere at a level that would prevent *dangerous* [our emphasis] anthropogenic interference with the climate system." The Framework Convention thus implies that, on the far side of some critical concentration level, climate change will be "dangerous," while on the near side of the threshold, climate change will be "safe" (though perhaps still undesirable). Rather than be linear and smooth, the Framework Convention warns that climate change may be "abrupt and catastrophic."

JEL-Code: C720, F510, H410, H870, Q540.

Keywords: climate change, prisoners' dilemma, catastrophe, negotiations, cooperation, uncertainty, experimental economics.

Scott Barrett
Lenfest-Earth Institute
School of International and Public Affairs
Columbia University
420 West 118th Street, Room 1427
USA - 10027 New York NY
sb3116@columbia.edu

Astrid Dannenberg
Lenfest-Earth Institute
Columbia University
420 West 118th Street, Room 1427
USA - 10027 New York NY
ad2901@columbia.edu

This Version: May 2013

What is the threshold for dangerous climate change? Climate negotiators first agreed on a value in the Copenhagen Accord, which recognizes "the scientific view that the increase in global temperature should be below 2 degrees Celsius." A year later, in Cancun, countries reaffirmed support for this goal, but added that the target might need to be strengthened to 1.5 °C. The threshold, it seems, is uncertain. A close reading of the scientific literature confirms this. There exists a range of values for the change in mean global temperature needed to "tip" critical geophysical systems (Lenton et al. 2008).

The temperature threshold is not the only uncertainty. The level of greenhouse gas concentrations needed to avoid any particular change in mean global temperature is also unknown (Roe and Baker 2007). Rockström et al. (2009: 473) combine both uncertainties to recommend a single target in terms of atmospheric CO_2 concentrations—350 parts per million by volume (ppmv). The value was chosen to preserve the polar ice sheets, and is derived from paleoclimatic evidence suggesting "a critical threshold between 350 and 550 ppmv." Rockström et al. essentially take a precautionary stance.

Even this may understate the uncertainties. Countries do not control atmospheric concentrations directly; they control emissions; and the relationship between emissions and concentrations is also uncertain. The stability of the carbon cycle itself cannot be taken for granted (Archer 2010).

All of this matters tremendously because recent research shows that uncertainty about the threshold for "dangerous" climate change can have a profound effect on international

cooperation. Theory predicts that countries can coordinate to avoid a "catastrophic" threshold so long as the threshold is known (and certain other conditions are satisfied), but that collective action collapses if the threshold is uncertain (Barrett 2013). Experimental evidence confirms this behavior (Barrett and Dannenberg 2012). When the threshold is certain, players coordinate to stave off "catastrophe." When the threshold is uncertain, they fail to cooperate so as to stay on the good side of the threshold.[1]

The literature on international environmental agreements has assumed that the underlying climate change game is a prisoners' dilemma. Until recently, however, this literature has ignored the possibility of thresholds. The key insight of the research reported above is that the climate change game may be a prisoners' dilemma for a different reason than assumed previously. The reason may not be that climate change is "gradual." The reason may be that climate change is "abrupt and catastrophic" but with an uncertain threshold.

Does the distinction matter? Analytical game theory predicts that behavior should be identical in both of these situations. Using our particular analytical model, free riding should cause countries to forego any abatement whether climate change is "gradual" and certain or "catastrophic" and uncertain. However, the *consequences* of free riding are worse when countries face the prospect of a looming "catastrophe." Moreover, numerous experiments have demonstrated that people tend to contribute more than predicted by analytical game theory, though less than is needed to supply an efficient amount of a public good (Ledyard 1995). It is as if pure self-interest pulls players in one direction (towards free riding), and group interest pulls them in another (towards full cooperation).

In short, people are conflicted; they are, after all, caught in a dilemma. This suggests that behavior may differ when countries try to cooperate to mitigate "gradual" climate change as opposed to "abrupt and catastrophic" climate change with an uncertain threshold since in the latter case they have much more to lose from the failure to cooperate.

In this paper we provide an experimental test of this hypothesis. Experiments give valuable insights into people's behavior when facing a collective action problem. Unlike analytical studies, they do not assume any particular preferences (for example, as regards selfishness or a willingness to take risks). Instead, they reveal how real people, having their own preferences, behave when facing a collective action problem. Our results confirm that the uncertain prospect of "catastrophe" increases abatement as compared to the prisoners' dilemma for "gradual" climate change. However, this result is merely a silver lining in an otherwise dark cloud, for our results also confirm that collective action fails to prevent "catastrophe." Given the scientific evidence for thresholds, negotiators were right to emphasize the need to avoid "dangerous" climate change early on. By doing so, our research suggests, global abatement probably increased. But due to scientific uncertainty about the location of the threshold—uncertainty that is substantially irreducible—knowledge of the existence of a threshold only helps in limiting "gradual" climate change. It won't help us to avert "catastrophe."

A simple analytical model

Our underlying game-theoretic model assumes a one-shot setting with N symmetric countries, each able to reduce emissions by up to $q^A_{\max}$ units using technology A and by up to $q^B_{\max}$ units using technology B. The per unit costs of reducing emissions by these two technologies are constant but different, with $c^A < c^B$. Technology A may be thought of as representing low-cost "ordinary abatement" and B as a high-cost technology for removing carbon dioxide from the atmosphere (Keith 2009). To understand why we include the latter technology, note that concentrations today are about 400 ppmv CO_2. If the aim were to limit concentrations to 350, as proposed by Rockström et al (2009) and others, we not only need to cut emissions substantially; we need to remove CO_2 from the atmosphere. Avoiding "catastrophe" may require dramatic action.

Let Q denote the reduction in emissions by all countries collectively using both technologies. Every unit of emission reduction gives each country a benefit in the amount b, the marginal benefit of avoiding "gradual" climate change. Assuming $c^B > bN > c^A > b$ gives the classical prisoners' dilemma in which self-interest and collective interest diverge. For these parameter values, self-interest impels each country to abate 0, whereas collectively all countries are better off if each abates $q^A_{\max}$ units using technology A and 0 units using technology B. Air capture is not worth doing in a world facing only "gradual" climate change.

Since climate thresholds can be related to cumulative emissions (Allen et al. 2009; Zickfeld et al. 2009), threshold avoidance can be expressed in terms of abatement relative to "business as usual." Denote the threshold by $\bar{Q}$, a parameter. Abatement short of this

value guarantees that the climate will tip "catastrophically," whereas abatement equal to or greater than this value preserves climate stability. Assume $N(q^A_{\max} + q^B_{\max}) > \bar{Q} > Nq^A_{\max}$. That is, avoidance of the threshold is technically feasible and requires using technology *B* in addition to *A*. Abatement short of $\bar{Q}$ results in "catastrophic" loss of value *X*. We restrict parameter values so that when countries cooperate fully they can do no better than to abate $\bar{Q}$ precisely, with technology A being fully deployed everywhere and technology B being used as a "top up" to make sure $Q = \bar{Q}$.

Acting independently, each country will maximize its own payoff, taking as given the abatement choices of other countries. We restrict parameter values so that, facing a certain threshold, there are two symmetric Nash equilibria in pure strategies.[2] In one, every country abates 0 and the threshold is exceeded. In the other, every country abates $q^A_{\max}$ using technology *A* and $\bar{Q}/N - q^A_{\max}$ using technology *B*, ensuring that the threshold is avoided, just.[3] By our restrictions, the latter equilibrium is universally preferred.[4] The game thus involves players coordinating to support this mutually preferred equilibrium.

With threshold uncertainty, $\bar{Q}$ is assumed to be distributed uniformly such that the probability of avoiding "catastrophe" is 0 for $Q < \bar{Q}_{\min}$, $\left(Q - \bar{Q}_{\min}\right)/\left(\bar{Q}_{\max} - \bar{Q}_{\min}\right)$ for $Q \in \left[\bar{Q}_{\min}, \bar{Q}_{\max}\right]$, and 1 for $Q > \bar{Q}_{\max}$. We assume $N\left(q^A_{\max} + q^B_{\max}\right) \geq \bar{Q}_{\max} > \bar{Q}_{\min} \geq Nq^A_{\max}$ and restrict parameters so that when countries cooperate fully they abate $\bar{Q}_{\max}$ collectively, eliminating threshold uncertainty, and when countries choose their abatement levels non-cooperatively, they do nothing to limit their emissions, making it

inevitable that the threshold will be crossed. For purposes of comparison, we assume that the expected value of the threshold is the same in the uncertainty case as in the certainty case.

Our analytical results and the parametrization of the experimental treatments are summarized in Table 1. The full cooperative abatement level is different for all three treatments. It is higher under *Certain Threshold* than under *No Threshold* because of the assumption that more abatement is needed to avoid "catastrophe" than is worth doing to limit "gradual" climate change. It is higher under *Uncertain Threshold* than under *Certain Threshold* because in this model (even assuming risk-neutral preferences) countries want to eliminate any chance of "catastrophe" (by assumption, the expected value for the threshold is the same under *Certain Threshold* and *Uncertain Threshold*). That is, it is in the collective interests of countries to act as if according to a precautionary principle.

INSERT TABLE 1 NEAR HERE

By contrast, the non-cooperative abatement level is predicted to be the same in the *No Threshold* and *Uncertain Threshold* treatments (zero abatement) but different for *Certain Threshold*. As explained before, for *Certain Threshold* there are two symmetric Nash equilibria, only one of which is efficient. With threshold certainty, abatement of greenhouse gases is a coordination game; and so long as the players can communicate,

there is strong reason to believe that countries will coordinate around the more efficient equilibrium.[5]

Our aim is to test these qualitative and quantitative predictions in the lab. In the next section we explain how our experiment was designed to allow us to do this.

Experimental design

At the start of every game, each subject was given "working capital" of €11, distributed between Accounts A (€1) and B (€10). Contributions to the public good consisted of poker chips (abatement) purchased from these accounts. Chips purchased from Account A cost €0.10 each ($c^A = 0.1$), and there were 10 chips ($q^A_{\max} = 10$). Chips paid for out of Account B cost €1.00 each ($c^B = 1$), and again there were 10 chips ($q^B_{\max} = 10$). Every subject was also given an endowment fund of €20, allocated to Account C. This fund could not be used to purchase chips; it was included only to ensure that no player could be left out of pocket. When the game was over, each subject received a payoff equal to the amount of money left in his or her three accounts, after making the following adjustment: Each subject was given €0.05 for every poker chip contributed by the group regardless of who had contributed that chip and from which account ($b = 0.05$). This treatment gives the classical prisoners' dilemma and is called *No Threshold*. Two more treatments included an additional adjustment: Each subject's payoff was reduced by €15 ($X = 15$) unless $\bar{Q}$ or more chips were contributed. In the *Certain Threshold* treatment, $\bar{Q}$ was set equal to 150 and in the *Uncertain Threshold* treatment, $\bar{Q}$ was assumed to be

distributed uniformly between 100 and 200. All parameter values are consistent with the expressions shown in Table 1.

The experimental sessions were conducted in a computer laboratory at the University of Magdeburg, Germany, using students recruited from the general student population (recruiting software Orsee; see Greiner 2004). In total, 300 students participated in the experiment, 100 per treatment: 10 groups × 10 students per group. In each session, subjects were seated randomly at linked computers (game software Ztree; see Fischbacher 2007). A set of written instructions including several numerical examples and control questions was handed out. The instructions involved a neutral frame for the experiment in order to avoid any potential biases the subjects may have regarding climate change. The control questions tested subjects' understanding of the game to ensure that they were aware of the available strategies and the implications of making different choices.

At the beginning of each session, subjects were assigned randomly to 10-person groups and played five practice rounds, with the membership of groups changing after each round. After a final reshuffling of members, each group played the game for real. To ensure anonymity, the members of each group were identified by the letters A through J. The game was played in stages; subjects first proposed a contribution target for the group and pledged an amount they each intended to contribute individually. It was common knowledge that these announcements were non-binding but would be communicated to the group. After being informed of everyone's proposals and pledges, subjects chose their

actual contributions in the second stage. The decisions in both stages were made simultaneously and independently. After the game, subjects were informed about everyone's decisions and asked to complete a short questionnaire, giving a picture of their reasoning, emotions, and motivation during the game. In the *Threshold Uncertainty* treatment, "Nature" chose the threshold in a third stage: a volunteer was invited to activate a computerized "spinning wheel" to determine the value for the threshold. This novel way of demonstrating a uniform distribution placed the minimum and maximum value of the threshold range (100 and 200) at the "ends" of the wheel at 12 o'clock.[6] Every subject was able to observe the wheel being spun and see where the arrow came to rest. At the end of each session, students were paid their earnings in cash.

Compared with the earlier literature, our experiment involves a threshold public goods game with no rebate (contributions above the threshold are not returned) and no refund (contributions are not returned if they fall short of the threshold) where the provision threshold is set to zero (*No Threshold*), or 150 (*Certain Threshold*), or is a random variable distributed uniformly between 100 and 200 (*Uncertain Threshold*).[7] Table 2 shows total contributions and individual payoffs corresponding to the three treatments.

INSERT TABLE 2 NEAR HERE

As shown in the table, compared to *No Threshold*, *Uncertain Threshold* increases the gap between the full cooperative and non-cooperative outcomes in terms of both contributions and payoffs. Previous papers have not made this same comparison, but they

have tested for the effect of increases in the gap between the full cooperative and non-cooperative payoffs in linear public goods games by increasing the marginal per capita return from the public good. There is robust evidence that an increase in the marginal return increases contributions (see Davis and Holt 1993, Ledyard 1995, and the literature cited therein). Therefore, although the theory predicts free riding behavior in both treatments, we may expect larger contributions in *Uncertain Threshold* than in *No Threshold*.

Results

Table 3 shows the summary statistics of the experimental data averaged across groups for each treatment.

INSERT TABLE 3 NEAR HERE

Look first at the proposals. In the *No Threshold* treatment, the mean proposal for the group target was 116. This is higher than the full cooperative level (100). However, the full cooperative level was proposed more often (by 63% of subjects) than any other value. The mean proposal for *Uncertain Threshold* was 166. This is lower than the full cooperative level (200). Once again, however, the full cooperative level was proposed more frequently than any other value (but by only 29% of subjects in this case). Finally, the mean proposal for *Certain Threshold* was 152. This is almost precisely equal to the full cooperative level (150), which was also by far the most frequent proposal (83% of

subjects, a remarkable degree of concordance). Taking groups as the unit of observation, a series of Mann-Whitney-Wilcoxon (MWW) tests shows that the differences in proposals between all three treatments are statistically significant (n = 20, p = .00 each; see Table 4).

Now consider the pledges. The mean pledge in the *No Threshold* treatment was 11. This is just a little over the full cooperative level, 10, which was the most frequently made pledge (announced by 65% of subjects). For the other prisoners' dilemma, *Uncertain Threshold*, the mean pledge was 16. This is below the most frequent pledge, which once again equals the full cooperative level (20, announced by 32% of subjects). Finally, in *Certain Threshold*, the mean pledge was equal to the most frequent pledge (15, announced by 74% of subjects), a value equal to the full cooperative level. The differences in pledges are significant between the *No Threshold* and the other two treatments (MWW test, n = 20, p = .00 each, see Table 4) and are weakly significant between *Certain Threshold* and *Uncertain Threshold* (n = 20, p = .06).

Lastly, look at the actual contributions. For *No Threshold*, the mean group contribution was 49, but the distribution of contributions varied widely, with most subjects contributing either zero (42%) or 10 (36%). Of course, the theory predicts that contributions should equal zero, but in this experiment chips contributed from Account A are very cheap. For *Uncertain Threshold*, the mean group contribution was 77. As in the *No Threshold* treatment, most subjects chose a contribution of either 10 (36%) or zero (30%). Behavior in the two prisoners' dilemma games was thus very similar. What

differed was the level of provision. Unlike the players in *No Threshold*, a few players in *Uncertain Threshold* threw in some of their expensive chips (25%), though only 2 out of 100 contributed all their expensive chips (see Figure 1). Finally, the mean group contribution for *Certain Threshold* was 151. This is just a hair over the predicted 150 (*T*-test, $n = 10$, $p = .72$), and in this case the most frequent individual contribution level equals the full cooperative level (15, chosen by 56% of subjects). The differences in contributions are significant between all three treatments (MWW test, $n = 20$, $p = .00$ each; see Table 4). The contributions in the two prisoners' dilemma games (*No Threshold* and *Uncertain Threshold*) are not only lower but also more erratic than those in the coordination game (*Certain Threshold*) (Levene test, $n = 20$, $p < .05$ each; see Tables 3 and 4). However, contributions in both *No Threshold* and *Uncertain Threshold* are significantly greater than the predicted zero (one-sided *T*-test, $n = 10$, $p = .00$ each).

INSERT TABLE 4 NEAR HERE

For *Certain Threshold* and *Uncertain Threshold*, the probability of "catastrophe" differs dramatically (MWW test, $n = 20$, $p = .00$). Eight out of 10 groups in *Certain Threshold* avoid "catastrophe." By contrast, "catastrophe" occurs in nine out of 10 cases for the *Uncertain Threshold* treatment, with the outlying group reducing the probability of "catastrophe" by just 7%.

Contributions in both of the prisoners' dilemma games are significantly lower than the proposals and pledges (Wilcoxon Signed-Rank test, $n = 10$, $p < .01$ each; see Figure 1).

By contrast, contributions in *Certain Threshold* are nearly equal to the amounts proposed and pledged.

INSERT FIGURE 1 NEAR HERE

Figure 1 shows the relationship between individual pledges (vertical axis) and individual contributions (horizontal axis). The correlation is positive and highly significant for *Certain Threshold* (Spearman's correlation test, $n = 100$, $rho = .38$, $p = .00$), while it is insignificant for both *No Threshold* ($n = 100$, $rho = -.03$, $p = .80$) and *Uncertain Threshold* ($n = 100$, $rho = .10$, $p = .34$). For *Certain Threshold,* almost all players (98%) contributed at least as much as they pledged, while only few players did so in *No Threshold* (33%) and *Uncertain Threshold* (18%).

Table 5 presents the responses to the ex-post questionnaire. Whenever questions are about the game rather than about general attitudes, the responses of the participants in the different treatments vary considerably, particularly as between the prisoners' dilemma and coordination games. For example, while fairness and trust are important driving forces for the contribution decisions in *Certain Threshold*, the coordination game, they are less relevant in *No Threshold* and *Uncertain Threshold*, the two prisoners' dilemmas. The proposals and in particular the pledges are perceived as being much more useful in *Certain Threshold* than in *No Threshold* and *Uncertain Threshold*. In coordination games, communication is particularly important.

The theory *assumes* that players are risk-neutral. Is this a reasonable assumption? As can be seen in Table 5 (see question 12), subjects' risk aversion does not differ significantly between treatments; the percentage of risk-averse subjects in each treatment is between 56% and 62%. Moreover, analysis of behavior within each treatment shows that there is no significant correlation between individual risk aversion and individual contributions or between the number of risk-averse members in a group and group contributions (Spearman's correlation test, $p > .10$ each). Thus, we cannot reject the hypothesis that risk aversion and behavior in the games are independent. People tend to play the same way in these games, irrespective of their preferences about risk.

Table 6 presents subjects' responses to the open questions about their motivation for making proposals, pledges, and contributions. The responses were classified according to key words and assigned to certain response categories. A large majority of the students playing *Certain Threshold* were motivated to state their proposal so as to maximize the joint group payoff (82%). By contrast, in the two prisoners' dilemma games, the motivation for making proposals was spread more evenly across three different responses—wanting to maximize joint payoffs, being realistic, and stimulating contributions by others. A large majority of the students playing in *Certain Threshold* used the pledge to signal truthfully their intended contribution and to create trust within the group (71%). By contrast, most subjects playing in *No Threshold* and *Uncertain Threshold* used the pledge to stimulate contributions by the other players (48% and 66%, respectively). As for the contribution decision, responses indicate that subjects in *Certain Threshold* were motivated either to contribute their fair share of the burden (56%) or to

compensate for potentially missing contributions (33%). Most subjects playing in the *No Threshold* and *Uncertain Threshold* treatments say that they chose their contributions because they wanted to maximize their own payoff (38% and 24%, respectively), because they distrusted the other players (20% and 30%), or because the chips were cheap (13% and 33%). Thus, there is a remarkable difference in the motivation and reasoning between the coordination and prisoners' dilemma games, indicating that the context of the games shapes people's beliefs and perceptions of appropriate behavior.

INSERT TABLE 5 NEAR HERE

INSERT TABLE 6 NEAR HERE

Conclusions

Theory predicts that behavior should be the same in the two prisoners' dilemma games—*No Threshold*, which corresponds to "gradual" climate change, and *Uncertain Threshold*, which corresponds to "dangerous" climate change with an uncertain threshold for "catastrophe." Our experimental results show that the motivations for the players were very similar in these games. However, the contributions varied. In both cases, the contributions exceeded the predicted amount, with the students playing the *Uncertain Threshold* prisoners' dilemma contributing more than the students playing the *No Threshold* prisoner's dilemma. This suggests that the framing of the negotiations around the need to avoid "dangerous" climate change has been advantageous. Countries may

reduce their emissions a bit more when facing this prospect than when they are ignorant about the prospect of "catastrophe." But our results also suggest that the warning about "dangerous" climate change will not suffice to cause countries to reduce their emissions by enough to avoid crossing the threshold.

The result that contributions are higher under *Uncertain Threshold* than under *No Threshold* confirms findings in other experimental settings, primarily linear public goods games. It is well established that people do not play a prisoners' dilemma precisely as predicted by analytical game theory; cooperation tends to be partial rather than completely absent. People are torn when playing a prisoners' dilemma. Their motives are mixed. Under *Uncertain Threshold*, playing only to please one's self-interest comes at a painful collective cost. And yet our results should offer little comfort. The players are not able to avoid the threshold. This is in complete contrast to how students play the game in which the threshold is certain. In this case, cooperation is enforced by "Mother Nature," which provides a sharp punishment for deviations from the full cooperative outcome. In the two prisoners' dilemma games, deviations from full cooperation are individually profitable for the players; to deter free riding, players must provide the punishment and enforcement themselves.

The policy implication is clear. The central challenge for a climate agreement is enforcement, and the Kyoto Protocol lacks an effective enforcement mechanism. Kyoto did not stop the United States from failing to participate. Nor did it create incentives for Canada to comply or to remain a party. Of course, many states have reduced their

emissions a little (Kyoto only aimed to reduce the emissions of Annex I countries by about five percent). However, this behavior is consistent with the players in our experiment contributing some of their cheap chips. The emission reductions needed to avoid the tipping points identified in the scientific literature require greater sacrifices. To avoid these levels, countries will have to hand in their expensive chips. Our research thus offers the following advice to negotiators: it is less important that countries agree on the collective target needed to avoid dangerous climate change than that they negotiate effective "strategic" mechanisms for enforcement (Barrett 2003). If "Mother Nature" doesn't provide enforcement, strategic mechanisms are needed to create the same conditions our experiment has revealed exist under a certain threshold.

Acknowledgments

We are grateful to James Rising for programming our "spinning wheel," and to the MaXLab team at Magdeburg University for allowing us to use their lab. This work was supported by the Swedish Research Council for the Environment, Agricultural Sciences and Spatial Planning through the program Human Cooperation to Manage Natural Resources.

Notes

[1] This same theory also predicts that uncertainty about the impact of crossing a threshold should have no effect on behavior (Barrett 2013)—another behavior confirmed by our experiment (Barrett and Dannenberg 2012). It is uncertainty in the threshold that matters.

[2] Of course, there are also many asymmetric Nash equilibria, but in our set up, contributions that are approximately symmetric are particularly focal.

[3] Starting from a situation in which every country abates $\bar{Q}/N$, should any one country reduce its abatement unilaterally by one unit, it will save c^B but lose $b+X$. Play $\bar{Q}/N$ is thus a Nash equilibrium so long as $X \geq c^B - b$.

[4] Avoiding "catastrophe" is mutually preferred so long as

$$b\bar{Q} - c^A q^A_{\max} - c^B\left(\bar{Q}/N - q^A_{\max}\right) \geq -X .$$

[5] If there is one thing climate negotiators can do it is communicate, which is why we include this possibility in our experiment. The experimental literature on communication has shown that restricted and anonymous communication, such as the proposals and pledges in our experiment, improves coordination but works much less reliably for cooperation (for reviews, see. Balliet 2010; Bicchieri and Lev-On 2007; Chaudhuri 2010; Crawford 1998; and Croson and Marks 2000). A previous threshold experiment by Milinski et al. (2008) found that the (certain) threshold was often crossed. However, this experiment did not allow communication. Tavoni et al. (2011) modified this experiment to show that communication significantly improves coordination.

[6] For details, see the Supplementary Information for Barrett and Dannenberg (2012).

[7] On threshold public goods experiments, see Bagnoli and Mckee (1991) and Croson and Marks (2000); on rebate rules in threshold public goods experiments, see Marks and Croson (1998).

References

Allen, M. R., D. J. Frame, C. Huntingford, C. D. Jones, J. A. Lowe, M. Meinshausen and N. Meinshausen (2009) Warming Caused by Cumulative Carbon Emissions Towards the Trillionth Tonne, *Nature,* 458, 1163-1166.

Archer, D. (2010) *The Global Carbon Cycle*. Princeton: Princeton University Press.

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*, Oxford: Oxford University Press.

Barrett, S. (2013) Climate Treaties and Approaching Catastrophes, *Journal of Environmental Economics and Management*, http://dx.doi.org/10.1016/j.jeem.2012.12.004.

Barrett, S. and A. Dannenberg (2012) Climate Negotiations Under Scientific Uncertainty, *Proceedings of the National Academy of Sciences,* 109 (43), 17372-17376.

Bicchieri, C. and A. Lev-On (2007) Computer-mediated Communication and Cooperation in Social Dilemmas: An Experimental Analysis, *Politics, Philosophy & Economics,* 6 (2), 139-168.

Bagnoli, M. and M. Mckee (1991) Voluntary Contribution Games: Efficient Private Provision of Public Goods, *Economic Inquiry*, 29 (2), 351–366.

Chaudhuri, A. (2011) Sustaining cooperation in laboratory public goods experiments: a selective survey of the literature, *Experimental Economics,* 14 (1), 47-83.

Crawford, V. (1998) A Survey of Experiments on Communication via Cheap Talk, *Journal of Economic Theory*, 78, 286-298.

Croson, R. T. A. and M. B. Marks (2000) Step Returns in Threshold Public Goods: A Meta- and Experimental Analysis, *Experimental Economics*, 2 (3), 239-259.

Davis, D. and C. Holt (1993) *Experimental Economics*, Princeton: Princeton University Press.

Fischbacher U. (2007) Z-Tree: Zurich Toolbox for Ready-made Economic Experiments, *Experimental Economics,* 10, 171-178.

Greiner B. (2004) An Online Recruitment System for Economic Experiments, In: Kremer, K. and V. Macho (eds.), *Forschung und wissenschaftliches Rechnen 2003*, GWDG Bericht 63, Göttingen, Ges. für Wiss. Datenverarbeitung, pp 79-93.

Keith, D.W. (2009) Why Capture CO_2 from the Atmosphere?, *Science,* 325 (5948), 1654-1655.

Ledyard, J. (1995) Public goods: a survey of experimental research, In: J.H. Kagel and A.E. Roth (eds.), *Handbook of Experimental Economics*, New Jersey: Princeton University Press, pp. 111-94.

Lenton, T. M., H. Held, E. Kriegler, J. W. Hal, W. Lucht, S. Rahmstorf and H. J. Schellnhuber (2008) Tipping Elements in the Earth's Climate System, *Proceedings of the National Academy of Sciences,* 105 (6), 1786-1793.

Marks, M. and R. T. A. Croson (1998) Alternative rebate rules in the provision of a threshold public good: An experimental investigation, *Journal of Public Economics,* 67 (2), 195-220.

Milinski M., R. D. Sommerfeld, H. J. Krambeck, F. A. Reed and J. Marotzke (2008) The Collective-Risk Social Dilemma and the Prevention of Simulated Dangerous Climate Change, *Proceedings of the National Academy of Sciences,* 105, 2291-2294.

Rockström, J., W. Steffen, K. Noone, Å. Persson, S. Chapin III, E. F. Lambin, T. M. Lenton, M. Scheffer, C. Folke, H. J. Schellnhuber, B. Nykvist, C. A. de Wit, T. Hughes, S. van der Leeuw, H. Rodhe, S. Sörlin, P. K. Snyder, R. Costanza, U. Svedin, M. Falkenmark, L. Karlberg, R. W. Corell, V. J. Fabry, J. Hansen, B. Walker, D. Liverman, K. Richardson, P. Crutzen and J. A. Foley (2009) A Safe Operating Safe for Humanity, *Nature* 461, 472-475.

Roe, G. H. and M. B. Baker (2007) Why is Climate Sensitivity So Unpredictable?, *Science,* 318, 629-632.

Tavoni, A., A. Dannenberg, G. Kallis and A. Löschel (2011) Inequality, Communication, and the Avoidance of Disastrous Climate Change in a Public Goods Game, *Proceedings of the National Academies,* 108 (29), 11825-11829.

Zickfeld, K., M. Eby, H. D. Matthews and A. J. Weaver (2009) Setting Cumulative Emissions Targets to Reduce the Risk of Dangerous Climate Change, *Proceedings of the National Academy of Sciences*, 106 (38), 16129-16134.

Table 1: Model summary and parametrization of the experimental treatments

Treatment	Game	Threshold	Full cooperation	Non-cooperation	Avoid "catastrophe"?
No Threshold	Prisoners' dilemma for "gradual" climate change	--	$Nq^{A}_{\max}$	0	--
			100	0	
Certain Threshold[*]	Coordination	$\bar{Q}$	$\bar{Q} > Nq^{A}_{\max}$	0, $\bar{Q}$	Yes[**]
		150	150	0, 150	
Uncertain Threshold[*]	Prisoners' dilemma for "dangerous" climate change	$\left[\bar{Q}_{\min}, \bar{Q}_{\max}\right]$	$\bar{Q}_{\max} > \bar{Q} > Nq^{A}_{\max}$	0	No
		[100, 200]	200	0	

[*]These treatments are taken from Barrett and Dannenberg (2012). [**]Assumes coordination on the efficient Nash equilibrium.

Table 2: Full cooperative and non-cooperative outcomes

Treatment	Total contributions			Individual payoffs		
	Full cooperation	Non-cooperation	Difference	Full cooperation	Non-cooperation	Difference
No Threshold	100	0	100	€35	€31	€4
Certain Threshold	150	150[*]	0	€32.5	€32.5[*]	€0
Uncertain Threshold	200	0	200	€30	€16	€14

[*]Assumes coordination on the efficient Nash equilibrium.

Table 3: Summary statistics

Treatment	Proposal		Pledge		Contribution		Group contribution
	Mean (Std dev)	Mode (%)	Mean (Std dev)	Mode (%)	Mean (Std dev)	Mode (%)	Min / max
No Threshold	115.8 (14.53)	100 (63%)	10.64 (1.18)	10 (65%)	4.9 (1.90)	0 (42%)	22 / 75
Certain Threshold	151.9 (1.57)	150 (83%)	14.7 (0.51)	15 (74%)	15.1 (0.77)	15 (56%)	136 / 159
Uncertain Threshold	166.3 (9.85)	200 (29%)	15.8 (1.69)	20 (32%)	7.7 (1.67)	10 (36%)	55 / 107

Mean and modal values for proposals, pledges, and contributions; standard deviations calculated with the group average taken as the unit of observation; percentages are shares of individuals per treatment. The rightmost column shows minimum and maximum group contributions for each treatment.

Table 4: Significance of treatment differences

	Proposal	Pledge	Contribution	Proposal	Pledge	Contribution
Certain Threshold	.0002 (.0001)	.0002 (.1642)	.0002 (.0044)			
Uncertain Threshold	.0002 (.1144)	.0002 (.2655)	.0041 (.5692)	.0002 (.0052)	.0638 (.0170)	.0002 (.0137)
	No Threshold			*Certain Threshold*		

p-values from a Mann-Whitney Wilcoxon rank-sum test of treatment differences in mean values; in parentheses *p*-values from a Levene test of treatment differences in variances.

Table 5: Responses to the ex-post questionnaire (percentages of subjects per treatment)

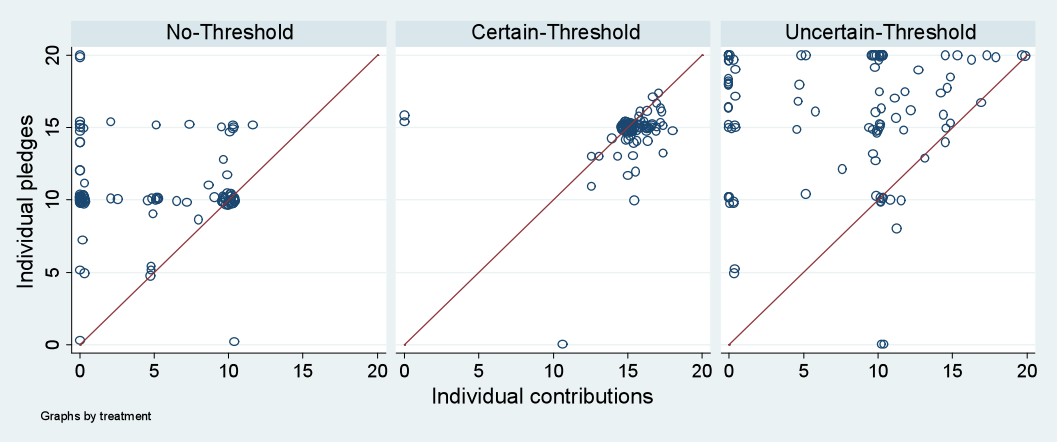
Question	Response	*No Threshold*	*Certain Threshold*	*Uncertain Threshold*
1) Were you generally satisfied with the game's outcome?	Very much	28	63	10
	Somewhat	48	18	31
	Little	16	5	26
	Not at all	8	14	33
2) Knowing how the game was played, with the benefit of hindsight, do you wish you had made a different contribution?	Very much	4	2	11
	Somewhat	15	19	17
	Little	17	27	22
	Not at all	64	52	50
3) Did fairness play a role for your contribution decision?	Very much	29	61	24
	Somewhat	23	16	10
	Little	17	11	21
	Not at all	31	12	45
4) Did trust play a role for your contribution decision?	Very much	23	58	18
	Somewhat	28	22	12
	Little	21	9	23
	Not at all	28	11	47
5) Do you agree with the statement that the exchange of proposals was helpful?	Very much	11	49	6
	Somewhat	37	27	28
	Little	33	13	34
	Not at all	19	11	32
6) Do you agree with the statement that the exchange of pledges was helpful?	Very much	13	68	10
	Somewhat	41	24	30
	Little	28	5	27
	Not at all	18	3	33
7) Generally speaking, do you trust other people?	Very much	24	25	21
	Somewhat	52	60	60
	Little	23	13	17
	Not at all	1	2	2
8) Generally speaking, do you agree with the statement that, if a person fails to keep his or her word, they deserve another chance?	Very much	40	24	41
	Somewhat	50	54	45
	Little	9	18	14
	Not at all	1	4	0
9) Generally speaking, do you try to keep your word?	Always	41	56	36
	Often	56	41	60
	Sometimes	2	1	4
	Rarely	1	1	0
	Never	0	1	0
10) Did you trust the other players to make the contributions they pledged?	Very much	8	47	10
	Somewhat	37	43	23
	Little	32	8	26
	Not at all	23	2	41
11) Knowing how the game was played, with the benefit of hindsight, do you feel, that some of the other players betrayed your trust in them?	Very much	7	10	16
	Somewhat	24	12	21
	Little	34	37	23
	Not at all	35	41	40
12) Please imagine the following situation in another unrelated experiment: You have an initial endowment of €40. There is a 50% possibility that you will lose your €40. However, you can avoid this loss by paying €20 up front. Would you rather pay this amount and get €20 for certain or would you rather accept the risk of losing the €40 with probability 50%?	€40 uncertain	21	15	25
	Indifferent	23	27	13
	€20 certain	56	58	62

Table 6: Responses to the ex-post questionnaire (open questions)

Question	Response	*No Threshold*	*Certain Threshold*	*Uncertain Threshold*
1) What was the most important reason for your proposal for the group contribution?	Joint payoff maximization	33	82	22
	Fairness	5	3	1
	Safety	2	8	0
	Stimulation of others' contributions	34	2	31
	Realistic target	19	0	39
	Other reason	7	5	7
2) What was the most important reason for your pledge for your own intended contribution?	Signaling of intended contribution/creation of trust	32	71	24
	Stimulation of others' contributions	48	17	66
	Safety	4	5	4
	Other reason	16	7	6
3) What was the most important reason for your contribution?	Fair share to reach target/own pledge	25	56	12
	Compensation of potentially missing contributions/safety	1	33	0
	Own payoff maximization	38	10	24
	Resignation/distrust	20	0	30
	Cheap chips/compromise between group and own interest	13	0	33
	Other reason	3	1	1

Percentages of subjects per treatment. These questions were posed as open questions; subjects' responses were classified by keyword search.

Figure 1: Correlation between pledges and contributions



A small noise (3%) has been inserted to make all data points visible.