Climate negotiations under scientific uncertainty

Scott Barrett[a,b,c,1] and Astrid Dannenberg[a,d]

[a]Earth Institute and [b]School of International and Public Affairs, Columbia University, New York, NY 10027; [c]Princeton Institute for International and Regional Studies, Princeton University, Princeton, NJ 08544; and [d]Department of Economics, University of Gothenburg, 405 30 Gothenburg, Sweden

How does uncertainty about "dangerous" climate change affect the prospects for international cooperation? Climate negotiations usually are depicted as a prisoners' dilemma game; collectively, countries are better off reducing their emissions, but self-interest impels them to keep on emitting. We provide experimental evidence, grounded in an analytical framework, showing that the fear of crossing a dangerous threshold can turn climate negotiations into a coordination game, making collective action to avoid a dangerous threshold virtually assured. These results are robust to uncertainty about the impact of crossing a threshold, but uncertainty about the location of the threshold turns the game back into a prisoners' dilemma, causing cooperation to collapse. Our research explains the paradox of why countries would agree to a collective goal, aimed at reducing the risk of catastrophe, but act as if they were blind to this risk.

Ever since the Framework Convention on Climate Change was adopted in 1992, negotiations over emission limits have been intertwined with efforts to identify a critical threshold for "dangerous anthropogenic interference with the climate system." A threshold finally was identified in the 2009 Copenhagen Accord: "the scientific view that the increase in global temperature should be below 2 degrees Celsius." However, the Copenhagen Accord relies on voluntary emission reductions to achieve this goal, and the amounts countries have pledged virtually guarantee that the 2 °C target will be missed (1). Identification of a threshold seems not to have helped the negotiations much at all.

Previous research suggests that this negative outcome is not inevitable but is largely a random occurrence, arising from a failure by negotiators to coordinate when the threshold is certain but the impact of crossing it is uncertain (2). Our research, which departs from the earlier literature in a number of ways (SI Literature), strongly questions this view. We provide experimental evidence suggesting that, if the threshold is known with certainty and the costs of avoiding it are low relative to the benefits, avoidance of the threshold is virtually assured whether or not the impact is uncertain, provided the negotiators can communicate (and if there is one thing negotiators can do it is communicate). Indeed, this finding may explain why the negotiations were framed around meeting a threshold and why negotiators wanted the threshold to be determined by "science" rather than by politics (only the former would be credible). Collective action fails, we show, because of uncertainty about the threshold. Far from being highly random, we show that failure is practically certain. Because the threshold is determined by Nature, and uncertainty about its value is substantially irreducible, our research suggests that negotiators should focus their attention on alternative strategies for collective action (3).

The scientific literature reveals not one but many scientific views about the temperature threshold for "dangerous" climate change (4–11), all of them uncertain. Even if a unique temperature threshold could be identified, countries can control only emissions directly, and the effect of emissions on temperature (mediated by the effect of emissions on atmospheric concentrations) is uncertain (12). Thresholds expressed in terms other than mean global temperature also are uncertain (13–16). One widely discussed paper identifies a unique "climate boundary" of 350 parts per million by volume (p.p.m.v.) atmospheric CO_2 "to ensure the continued existence of the large polar ice sheets," for

which "there is a critical threshold between 350 and 550 p.p.m.v." (16). Our model can be interpreted as representing threshold uncertainty in this same way. Using the above reference values, our model suggests that countries can recognize that it is best to limit concentrations to 350 p.p.m.v. but still be compelled in this prisoners' dilemma to propose a higher target, to pledge less than is needed to meet this target, and then to contribute less than they pledged, with the consequence that concentrations ultimately exceed 550 p.p.m.v.

Although our paper was motivated by the climate problem, the participants in our experiment were not told of this motivation, making our results equally applicable to other situations in which collective action is needed to avoid a dangerous threshold. Examples range from the cascading effect of adding space debris beyond a critical level, rendering a key orbit unusable (17), to thresholds in antibiotic use, causing a disease to become drug resistant (18). Another example is the negotiation of fishery quotas—a routine task for the world's 17 regional fishery management organizations. For many species, there exists a critical minimum population level, but with unknown value. Making matters worse, fish stocks cannot be observed directly, and catch-per-unit-of-effort may fail to signal an impending crash, perhaps because of technological change (19) or the tendency of some species of fish to aggregate (20). When combined, these conditions can create a true tragedy of the commons. In all these situations, as in our game, countries have a collective incentive to avoid the far-reaching consequences of exceeding a threshold but also face individual incentives to free ride because of the inherent uncertainty about the location of the threshold.

Our underlying game-theoretic model assumes that there are N symmetric countries, each able to reduce emissions by up to q_{max}^A units using technology A and by up to q_{max}^B units using technology B. The per-unit cost of reducing emissions by these means are constant but different, with $c^A < c^B$. We can think of A as representing low-cost "ordinary abatement" and B as a high-cost technology for removing carbon dioxide from the atmosphere (21). Q denotes the total reduction in emissions by all countries using both technologies. Every unit of emission reduction gives each country a benefit, b, the marginal benefit of avoiding "gradual" climate change. Assuming $c^B > bN > c^A > b$ gives the classical prisoners' dilemma. For these parameter values, self-interest impels each country to abate 0, whereas collectively all countries are better off if each abates q_{max}^A units using technology A and 0 units using technology B.

Because climate thresholds can be related to cumulative emissions (22, 23), threshold avoidance can be expressed in terms of abatement from business as usual. Denote the threshold by $\overline{Q}$ and assume $N(q_{max}^A + q_{max}^B) > \overline{Q} > Nq_{max}^A$. That is, avoidance of the threshold is technically feasible and requires using B in addition to A (air capture is needed to reduce concentrations
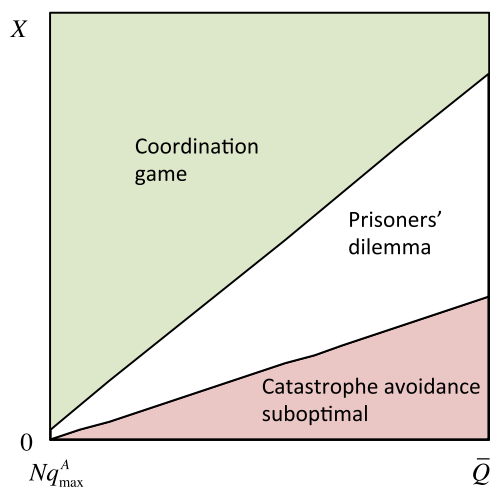
Fig. 1. Certainty model. Red area shows values for X and $\overline{Q}$ for which countries are collectively better off *not* avoiding catastrophe; here, $X < (c^B - bN)(\overline{Q}/N - q_{max}^A)$. In the green area, catastrophe avoidance is a coordination game; here, $X \geq (c^B - b)\overline{Q}/N - (c^B - c^A)q_{max}^A$. In the white area, avoiding catastrophe is a prisoners' dilemma; here, if all other countries play $\overline{Q}/N$, each country prefers to abate 0. With certainty, a prisoners' dilemma arises only if $b > 0$.

from today's level to 350 p.p.m.v.). Abatement short of $\overline{Q}$ 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 $\overline{Q}$ precisely (Fig. 1), with technology A being fully deployed everywhere and technology B being used as a "top up" to make sure $Q = \overline{Q}$.

Acting independently, each country will maximize its own payoff, taking as given the abatement choices of other countries. We restrict parameter values so that there are two symmetric Nash equilibria in pure strategies. In one, every country abates 0, and the threshold is exceeded. In the other, every country abates q_{max}^A using technology A and $\overline{Q}/N - q_{max}^A$ using technology B, ensuring that the threshold is narrowly avoided. By our restrictions, the latter equilibrium is universally preferred. The game thus involves players coordinating to support this mutually preferred equilibrium (Fig. 1).

With threshold uncertainty, $\overline{Q}$ is assumed to be distributed uniformly so that the probability of avoiding catastrophe is 0 for $Q < \overline{Q}_{min}$, $(Q - \overline{Q}_{min})/(\overline{Q}_{max} - \overline{Q}_{min})$ for $Q \in [\overline{Q}_{min}, \overline{Q}_{max}]$, and 1 for $Q > \overline{Q}_{max}$. We assume $N(q_{max}^A + q_{max}^B) \geq \overline{Q}_{max} > \overline{Q}_{min} \geq Nq_{max}^A$ and restrict parameters so that when countries cooperate fully they abate $\overline{Q}_{max}$ collectively, eliminating threshold uncertainty, and when countries choose their abatement levels noncooperatively, they do nothing to limit their emissions, making it inevitable that the threshold will be crossed. Our experiment also assumes a uniform distribution for impacts, which means X must be replaced by its expected value in our analytical model.

Our experiment involved 400 participants (*Materials and Methods* and *SI Materials and Methods*): 10 games per treatment × 4 treatments × 10 players per game. At the start of each game, every 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 ($q_{max}^A = 10$). Chips paid for out of Account B cost €1.00 each ($c^B = 1$), and again there were 10 chips ($q_{max}^B = 10$). Every subject also was 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.

After the game was played, each subject received a payoff equal to the amount of money left in his or her three accounts, after making two further adjustments. First, each subject was given €0.05 for every poker chip contributed by the group ($b = 0.05$). Second, each subject's payoff was reduced by an amount X unless $\overline{Q}$ or more chips were contributed. In the *Certainty* treatment, $X = €15$ and $\overline{Q} = 150$. Under *Impact Uncertainty*, X was distributed uniformly between €10 and €20. Under *Threshold Uncertainty*, $\overline{Q}$ was distributed uniformly between 100 and 200. In the *Impact-and-Threshold Uncertainty* treatment, X and $\overline{Q}$ were both distributed uniformly as above.

The game was played in stages. In the communication stage, every subject pledged an amount he or she intended to contribute individually and also proposed a contribution target for the group. It was common knowledge that proposals and pledges were nonbinding. Once every member of a group had made these choices, all members were informed about these values. In the contributions stage, subjects chose their actual contributions. Then the players were informed about everyone's individual and collective contributions.

For the uncertainty treatments, "Nature" chose the impact and/or the threshold in a third stage. Probabilities can be difficult for people to comprehend and so must be communicated with care (24). In our game, a volunteer was invited to activate a computerized "spinning wheel," with the "ends" of the wheel at 12 o'clock representing the minimum and maximum values of the range [(€10, €20) for X and (100, 200) for $\overline{Q}$]. Every subject was able to observe the wheel being spun and see where the arrow came to rest, determining the value for the impact and/or the threshold. After completing a follow-up questionnaire, participants were paid their earnings in cash. Answers to the survey indicate that the players understood the games and the probabilities determined by the spinning wheel (*SI Materials and Methods*).

Results

Our main hypotheses are that catastrophe will be avoided in the *Certainty* and *Impact Uncertainty* treatments but not in the *Threshold Uncertainty* and *Impact-and-Threshold Uncertainty* treatments. Our main results strongly support both hypotheses (Fig. 2). The difference in the frequency of catastrophe between *Certainty* and *Impact Uncertainty*, on the one hand, and *Threshold Uncertainty* and *Impact-and-Threshold Uncertainty,* on the other,
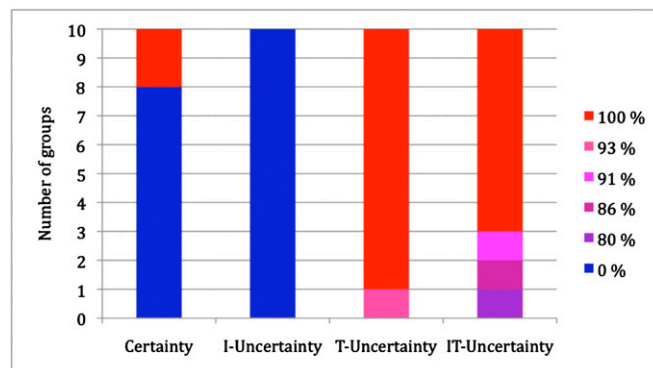

Fig. 2. Probability of catastrophe by treatment. Catastrophe was avoided 8 of 10 times in the *Certainty* treatment and 10 of 10 times under *Impact Uncertainty* (I-Uncertainty). In contrast, the probability of catastrophe was reduced below 100% (to 93%) by only 1 of 10 groups under *Threshold Uncertainty* (T-Uncertainty) and by only 3 of 10 groups (to 91, 86, and 80%, respectively) under *Impact-and-Threshold Uncertainty* (IT-Uncertainty). In the four cases where the probability of catastrophe was reduced below 100%, the spinning wheel determined that the threshold was crossed every time.

Table 1. Summary statistics: Mean values across groups per treatment

Treatment	Mean proposal	Mean pledge	Mean group contribution	Range of group contribution
Certainty	151.9 (1.57)	14.7 (0.51)	150.9 (7.69)	136–159
Impact Uncertainty	149.1 (4.98)	14.4 (0.80)	155.5 (2.92)	152–161
Threshold Uncertainty	166.3 (9.85)	15.8 (1.69)	77.2 (16.67)	55–107
Impact-and-Threshold Uncertainty	167.0 (10.40)	15.5 (2.07)	79.9 (26.90)	44–120

SDs are given in parentheses. Under threshold uncertainty, players propose that *more* be contributed, compared with the threshold certainty treatments, but end up contributing *less*. Variability in proposals, pledges, and especially contributions also is greater for the threshold uncertainty treatments.

is statistically significant (Fisher's exact test, $n = 20$, $P < 0.05$ each). In the two threshold certainty treatments, catastrophe was avoided 18 of 20 times. (In each of the two cases in which catastrophe was not avoided, the reason was a sharp deviation from the pledged and expected behavior of a single individual; see below.) In the two threshold uncertainty treatments, catastrophe occurred with certainty in 16 of 20 cases and with a probability of at least 80% in the other four cases.

As predicted, group contributions are significantly lower in the treatments with threshold uncertainty than in those without threshold uncertainty (Table 1, Mann–Whitney–Wilcoxon test, $n = 20$, $P < 0.05$ each; *SI Results*). The former also exhibit greater variability (Levene test, $n = 20$, $P < 0.05$ each). There are no statistically significant differences within these pairs of treatments. That is, impact uncertainty has no significant effect on collective action.

In both the *Certainty* and *Impact Uncertainty* treatments, group contributions are relatively close to the predicted 150. In both treatments the most frequent individual contribution is 15, the obvious focal point (25). Fifty-six percent of subjects chose this contribution level in *Certainty*. Fifty percent did so in *Impact Uncertainty*.

The prediction of zero contributions in the two threshold uncertainty treatments, on the other hand, is clearly rejected (one-sided *t* test, $n = 10$, $P = 0.00$ each). Zero individual contributions were common (30% in *Threshold Uncertainty* and 32% in *Impact-and-Threshold Uncertainty*), but contributions of 10

were slightly more common (36% in *Threshold Uncertainty*, 39% in *Impact-and-Threshold Uncertainty*). These subjects contributed from their low-cost account to lessen the well-known conflict between collective and individual interests (26–28).

Communication is the essence of negotiation, and it is striking how the players used their proposals and pledges differently depending on threshold uncertainty. When the threshold was known, players communicated so as to coordinate to the threshold. When the threshold was unknown, communication was more strategic. Mean proposals for the *Certainty* and *Impact Uncertainty* treatments are very close to 150 (Table 1), with 83% of subjects in *Certainty* and 94% in *Impact Uncertainty* proposing precisely this amount. Mean proposals in *Threshold Uncertainty* and *Impact-and-Threshold Uncertainty* were significantly larger (Mann–Whitney–Wilcoxon test, $n = 20$, $P < 0.05$ each), with 29% of subjects in *Threshold Uncertainty* and 35% in *Impact-and-Threshold Uncertainty* proposing 200. Why did not more participants propose the collectively optimal 200? Answers to questions in our follow-up questionnaire provide a strong clue. Participants perceived their proposals as serving to motivate other students to contribute; they thought that a proposal below 200 was more credible and so was more likely to stimulate contributions by others.

Fig. 3 shows the relationship between pledges and actual contributions. In *Certainty* and *Impact Uncertainty*, almost all players (98% in both treatments) contributed at least as much as they pledged. Two of 200 contributed substantially less than they pledged, causing the two breakdowns in collective action in the
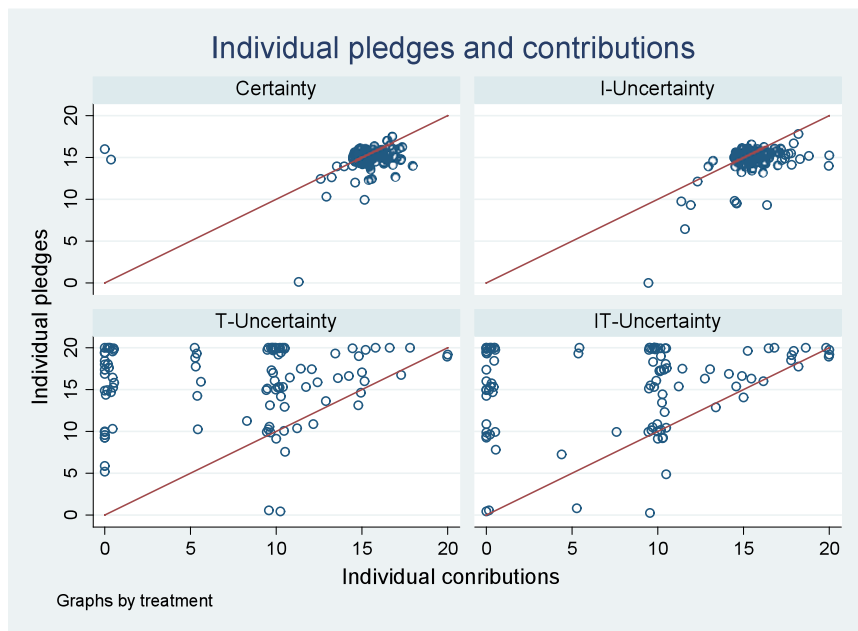

Individual pledges and contributions

Fig. 3. Pledges and actual contributions by treatment. In the *Certainty* and *Impact Uncertainty* treatments, pledges and contributions are tightly bunched, with contributions usually exceeding pledges. In the *Threshold* and *Impact-and-Threshold Uncertainty* treatments, values vary widely, with contributions usually falling far short of pledges. A small noise (3%) has been inserted to make all data points visible.

Certainty treatment. By contrast, in the *Threshold Uncertainty* and *Impact-and-Threshold Uncertainty* treatments, most (82 and 75%, respectively) contributed less than they pledged, indicating that pledges, like proposals, were used strategically.

Our follow-up questionnaire revealed that the reason for these differences had to do with the context in which decisions were made. Fairness and trust were more important considerations for the coordination games than for the prisoners' dilemmas. Players were more trusting in the *Certainty* and *Impact Uncertainty* treatments because each recognized that the others had a strong incentive to be trustworthy in these situations.

A final observation concerns attitudes toward risk, which can play a crucial role in the analysis of collective best outcomes (29). Our theory assumes that people are risk-neutral. Our questionnaire reveals that a majority of subjects are risk averse, but statistical analysis shows that whether a person is risk averse has no discernable influence on behavior (*SI Results*). Once again, the context of these games seems to shape how people behave.

Discussion

There is universal agreement among countries that global emissions should be limited so as to prevent "dangerous interference" with the climate system. Our research strongly suggests that if a threshold for catastrophic climate change could be identified with certainty, free-riding behavior would be disciplined; countries very likely would propose a collective target certain to avoid catastrophe, would pledge to contribute their fair share to the global effort, and would act so as to fulfill their promises. Scientists have endeavored to support this negotiation strategy by identifying a "red line" for collective action, but thresholds for "abrupt and catastrophic" climate change are inherently uncertain. Our research suggests that, under these circumstances, countries are very likely to propose to do less collectively than is needed to avert catastrophe, pledge to contribute less than their fair share of the amount proposed, and end up contributing even less than their pledge. The climate change game is a prisoners' dilemma, but not for the reasons usually given. What makes it a prisoners' dilemma is not just the need for collective action but uncertainty about the threshold for dangerous climate change.

Our analysis is consistent with how the climate negotiations have played out so far. Concern about climate thresholds has reinforced the need to limit emissions so as to reduce, if not eliminate, the risk of dangerous interference, without having any noticeable effect on how countries behave. As in our experiment, countries have pledged to do less than is needed to meet their stated collective goals. We will not know until 2020 if the Copenhagen Accord pledges will be met, but if our experimental results are a reliable guide, countries may end up emitting more than they pledged—with potentially profound and possibly irreversible consequences.

Our research thus underscores the need to pursue alternative negotiation strategies for transforming the prisoners' dilemma.

Collective action can succeed, we have shown, when the underlying prisoners' dilemma game is transformed into a coordination game. Although threshold uncertainty spoils this transformation, previous research shows that strategic treaty design can bring about a similar transformation. One way is by the use of trade restrictions against nonparticipating countries. If the loss from the trade restrictions exceeds the gains from free riding, every country will want to participate in a treaty, so long as each is assured that others will participate; this is how the Montreal Protocol enforced restrictions on the production and consumption of chlorofluorocarbons to protect the ozone layer (3). Another way to make abatement a coordination game is by the use of technology standards when these exhibit network externalities—that is, when the returns to each country of adopting a standard increase with the number of other countries that adopt the standard (30); this is how the MARPOL treaty limited releases of oil into the sea by tankers (3). Climate change is a more complex challenge, but our research suggests that strategies like these will be more successful than relying exclusively on the fear of dangerous climate change.

Materials and Methods

The experimental sessions were held in a computer laboratory at the University of Magdeburg, Germany, using students recruited from the general student population. In total, 400 students participated in the experiment, 100 per treatment. At the beginning of a session, subjects were seated at computers, which were linked to enable structured communication during the game (see *SI Materials and Methods* for further details and software). Written instructions, including several numerical examples and control questions, were handed out. The control questions tested the subjects' understanding of the game to ensure that they were aware of the implications of making different choices. Subjects then were assigned randomly to 10-person groups and played five practice rounds, with the membership of each group changing after each round. After a final reshuffling of members, each group played the game itself. To ensure anonymity, the members of each group were identified by the letters A through J. Subjects first announced a contribution target for the group and an amount they intended to contribute themselves. After being informed about everyone's proposals and pledges, subjects chose their actual contributions. The decisions in both stages were made simultaneously and independently. Players were informed about all the decisions at the end of the game. They also were informed about individual expected payoffs contingent on the probability of the loss and the expected value of the loss. After the game, subjects were asked to complete a short questionnaire, giving a picture of their reasoning, emotions, and motivation during the game. Then they were paid their earnings in cash.

1. Rogelj J, et al. (2010) Copenhagen Accord pledges are paltry. *Nature* 464(7292): 1126–1128.
2. Milinski M, Sommerfeld RD, Krambeck H-J, Reed FA, Marotzke J (2008) The collective-risk social dilemma and the prevention of simulated dangerous climate change. *Proc Natl Acad Sci USA* 105(7):2291–2294.
3. Barrett S (2003) *Environment and Statecraft* (Oxford Univ Press, Oxford).
4. Hansen J, et al. (2007) Dangerous human-made interference with climate: A GISS modelE study. *Atmos Chem Phys* 7:2287–2312.
5. Hansen J, et al. (2008) Target atmospheric CO2: Where should humanity aim? *Open Atmospheric Science Journal* 2:217–231.
6. Oppenheimer M, Alley RB (2004) The West Antarctic ice sheet and long term climate policy. *Climatic Change* 64:1–10.
7. Oppenheimer M, Alley RB (2005) Ice sheets, global warming, and article 2 of the UNFCCC - An Editorial Essay. *Climatic Change* 68:257–267.
8. Smith JB, et al. (2001) *Climate Change 2001: Impacts, Adaptation, and Vulnerability*, eds McCarthy J, Canziana O, Leary N, Dokken D, White K (Cambridge Univ Press, New York), pp 913–967.
9. Smith JB, et al. (2009) Assessing dangerous climate change through an update of the Intergovernmental Panel on Climate Change (IPCC) "reasons for concern" *Proc Natl Acad Sci USA* 106(11):4133–4137.
10. O'Neill BC, Oppenheimer M (2002) Climate change. Dangerous climate impacts and the Kyoto Protocol. *Science* 296(5575):1971–1972.
11. Lenton TM, et al. (2008) Tipping elements in the Earth's climate system. *Proc Natl Acad Sci USA* 105(6):1786–1793.
12. Roe GH, Baker MB (2007) Why is climate sensitivity so unpredictable? *Science* 318 (5850):629–632.
13. Petschel-Held G, Schellnhuber HJ, Bruckner T, Hasselmann K, Toth F (1999) The tolerable windows approach. An inverse assessment of climate change. *Climatic Change* 41:303–331.
14. Mastrandrea MD, Schneider SH (2004) Probabilistic integrated assessment of "dangerous" climate change. *Science* 304(5670):571–575.
15. Lenton TM (2011) Beyond 2°C: Redefining dangerous climate change for physical systems. *Wiley Interdiscip Rev Clim Change* 2:451–461.
16. Rockström J, et al. (2009) A safe operating space for humanity. *Nature* 461(7263): 472–475.

ENVIRONMENTAL SCIENCES

SUSTAINABILITY SCIENCE

17. Kessler DJ, Cour-Palais BG (1978) Frequency of artificial satellites: The creation of a debris belt. *Journal of Geophysical Research* 83:2637–2646.

18. Austin DJ, Kristinsson KG, Anderson RM (1999) The relationship between the volume of antimicrobial consumption in human communities and the frequency of resistance. *Proc Natl Acad Sci USA* 96(3):1152–1156.

19. Myers RA, Hutchings JA, Barrowman NJ (1997) Why do fish stocks collapse? The example of cod in Atlantic Canada. *Ecol Appl* 7:91–106.

20. Clark CW, Mangel M (1979) Aggregation and fishery dynamics: A theoretical study of schooling and the purse seine tuna fisheries. *Fishery Bulletin* 77:317–337.

21. Keith DW (2009) Why capture CO_2 from the atmosphere? *Science* 325(5948):1654–1655.

22. Allen MR, et al. (2009) Warming caused by cumulative carbon emissions towards the trillionth tonne. *Nature* 458(7242):1163–1166.

23. Zickfeld K, Eby M, Matthews HD, Weaver AJ (2009) Setting cumulative emissions targets to reduce the risk of dangerous climate change. *Proc Natl Acad Sci USA* 106 (38):16129–16134.

24. Spiegelhalter D, Pearson M, Short I (2011) Visualizing uncertainty about the future. *Science* 333(6048):1393–1400.

25. Schelling TC (1960) *The Strategy of Conflict* (Harvard Univ Press, Cambridge MA).

26. Ledyard JO (1995) *The Handbook of Experimental Economics*, eds Kagel JH, Roth AE (Princeton University Press, Princeton), pp 111–194.

27. Ostrom E (1998) A behavioral approach to the rational choice theory of collective action. *American Political Science Review* 92:1–22.

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

29. Weitzman ML (2009) On modeling and interpreting the economics of catastrophic climate change. *Rev Econ Stat* 91:1–19.

30. Barrett S (2006) Climate treaties and 'breakthrough' technologies. *Am Econ Rev* 96:22–25.

Barrett and Dannenberg