LETTERS
PUBLISHED ONLINE: 8 DECEMBER 2013 | DOI: 10.1038/NCLIMATE2059
nature climate change

Sensitivity of collective action to uncertainty about climate tipping points

Scott Barrett[1,2]* and Astrid Dannenberg[3,4]

Despite more than two decades of diplomatic effort, concentrations of greenhouse gases continue to trend upwards, creating the risk that we may someday cross a threshold for 'dangerous' climate change[1–3]. Although climate thresholds are very uncertain, new research is trying to devise 'early warning signals' of an approaching tipping point[4–11]. This research offers a tantalizing promise: whereas collective action fails when threshold uncertainty is large, reductions in this uncertainty may bring about the behavioural change needed to avert a climate 'catastrophe'[5]. Here we present the results of an experiment, rooted in a game-theoretic model, showing that behaviour differs markedly either side of a dividing line for threshold uncertainty. On one side of the dividing line, where threshold uncertainty is relatively large, free riding proves irresistible and trust illusive, making it virtually inevitable that the tipping point will be crossed. On the other side, where threshold uncertainty is small, the incentive to coordinate is strong and trust more robust, often leading the players to avoid crossing the tipping point. Our results show that uncertainty must be reduced to this 'good' side of the dividing line to stimulate the behavioural shift needed to avoid 'dangerous' climate change.

Our approach can be applied to a variety of situations, from the collapse of a fishery to sudden transitions of ecological[6,12–15] or other complex systems[6,16], but our focus in this paper is on averting 'dangerous' climate change[1–11], perhaps the greatest challenge for collective action ever. We begin by presenting a game-theoretic model of countries' decisions to limit their greenhouse gas emissions and avoid crossing a critical tipping point. In our model, avoiding the tipping point is feasible and collectively optimal, but individual optimization by countries sustains this desirable outcome only when uncertainty about the location of the threshold is sufficiently small; when uncertainty is larger, our model predicts that individual behaviour, motivated by self-interest, will push countries over the tipping point, resulting in catastrophe. Our main contribution is to test this prediction of a behavioural regime shift in the laboratory.

Our game-theoretic model assumes that there are N identical countries, each able to reduce emissions by up to $q^{\mathrm{A}}_{\max}$ units using technology A and by up to $q^{\mathrm{B}}_{\max}$ units using technology B. The per-unit costs of reducing emissions using these technologies are constant but different, with $c^{\mathrm{A}} < c^{\mathrm{B}}$. Technology A can be thought of as low-cost 'ordinary abatement', and B as a high-cost technology for removing carbon dioxide from the atmosphere[17].

Let Q denote the total reduction in emissions by all countries using both technologies, and let b represent the marginal benefit to an individual country of avoiding 'gradual' climate change. Assuming $c^{\mathrm{B}} > bN > c^{\mathrm{A}} > b$ gives the classical prisoners' dilemma. For these parameter values, self-interest impels each country to abate zero, whereas collectively all countries are better off when each abates $q^{\mathrm{A}}_{\max}$ units using technology A. Technology B would never be used to address 'gradual' climate change in this model, but is essential for avoiding the tipping point.

As climate thresholds can be related to cumulative emissions[18,19], threshold avoidance can be expressed in terms of abatement from business as usual. Denote the threshold by $\overline{Q}$, and assume that $\overline{Q}$ is a random variable distributed uniformly such that the probability of avoiding the threshold 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 that avoidance of the threshold is technically feasible but requires using technology B in addition to A. Abatement short of $\overline{Q}$ results in loss of value X. Theory[20] and experimental evidence[21] suggest that uncertainty about the impact of crossing the threshold, X, should not affect behaviour and so we assume that the value of X is certain.

We next solve two different optimization problems (see Methods and Supplementary Methodological Details). We first show that all countries collectively will want to abate $\overline{Q}_{\max}$ so long as $X \geq (c^{\mathrm{B}} - bN)(\overline{Q}_{\max} - Nq^{\mathrm{A}}_{\max})/N$. In this paper we assume that this condition is always satisfied, making the consequences of crossing the tipping point truly catastrophic.

We next show that if every other country abates $\overline{Q}_{\max}/N$, each will want to abate $\overline{Q}_{\max}/N$, making the avoidance of catastrophe a Nash equilibrium, provided $X \geq (c^{\mathrm{B}} - b)(\overline{Q}_{\max} - \overline{Q}_{\min})$. Of course, in this game zero abatement is also a Nash equilibrium. Hence, when this second condition holds, the collective action problem is for countries to coordinate their abatement so as to sustain the 'safe' Nash equilibrium. When this condition does not hold, the players are trapped in a prisoners' dilemma; they would rather stay on the safe side of the tipping point, but free-rider incentives draw them inexorably towards the unique but catastrophic equilibrium.

Rearranging this second condition, we can define $\phi = X/(c^{\mathrm{B}} - b)$ to be the dividing line for the range of threshold uncertainty. Assuming that all countries abate the same amount, our model thus predicts that countries will play $q^{\mathrm{A}}_i = q^{\mathrm{A}}_{\max}$, $q^{\mathrm{B}}_i = \overline{Q}_{\max}/N - q^{\mathrm{A}}_{\max}$ to the left of the dividing line, where $\overline{Q}_{\max} - \overline{Q}_{\min} \leq \phi$, and $q^{\mathrm{A}}_i = q^{\mathrm{B}}_i = 0$ to the right of the dividing line, where $\overline{Q}_{\max} - \overline{Q}_{\min} > \phi$. (Note that, in our experiment, because abatement is expressed in discrete units, the efficient outcome may not be exactly symmetric; see Supplementary Methodological Details.)

In the experiment, the game is played by groups of ten players. At the start of each game, every participant was given €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

[1]Earth Institute and School of International and Public Affairs, Columbia University, New York, New York 10027, USA, [2]Princeton Institute for International and Regional Studies, Princeton University, Princeton, New Jersey 08544, USA, [3]Earth Institute, Columbia University, New York, New York 10027, USA, [4]University of Gothenburg, Gothenburg 405 30, Sweden. *e-mail: sb3116@columbia.edu

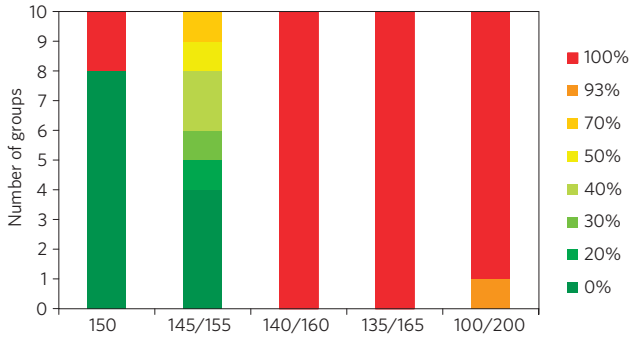

Figure 1 | Probability of catastrophe by treatment. In 150, catastrophe is avoided eight out of ten times. In 145/155, catastrophe is avoided four out of ten times with probability 100% and in the other six cases with probability between 30 and 80%. In 140/160 and 135/165, catastrophe is never avoided. In 100/200, catastrophe occurs nine out of ten times with probability 100% and once with probability 93%.

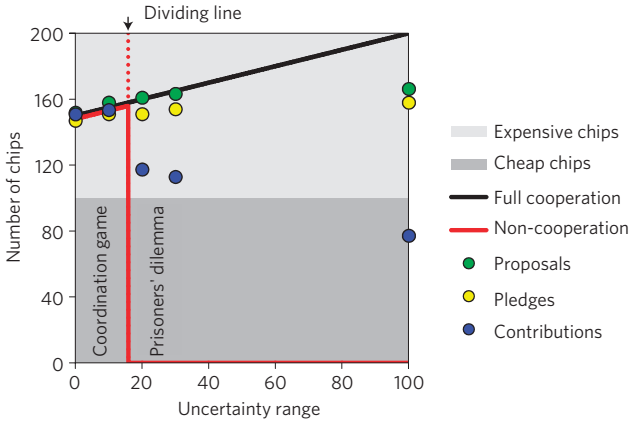

Figure 2 | Treatment means versus predicted values. Mean contributions are consistent with the predicted values to the left of the dividing line. To the right of the dividing line, mean contributions lie between the full cooperative and the predicted (non-cooperative) values. Mean proposals and mean pledges match the full cooperative values to the left of the dividing line; to the right, a wedge opens up between these values as the uncertainty range widens.

($c^{A} = 0.1$), and there were ten chips ($q^{A}_{max} = 10$). Chips paid for out of Account B cost €1.00 each ($c^{B} = 1$), and again there were ten chips ($q^{A}_{max} = 10$). Participants were also given an 'endowment fund' of €20, allocated to Account C. This fund could not be used to contribute to the public good; it was included only to ensure that no player could be left out of pocket. At the conclusion of the game, each participant received a payoff equal to the amount of money left in his or her accounts, after making two adjustments: first, each subject was given €0.05 for every poker chip contributed by the group ($b = 0.05$) regardless of who had contributed the chips and from which account they had been taken. Second, each subject's payoff was reduced by €15 ($X = 15$) unless $\overline{Q}$ or more chips were contributed by the group. In the baseline treatment, the threshold was certain and set equal to $\overline{Q} = 150$ (chips). In the other treatments, the threshold was a random variable, distributed uniformly over a range of values: 145/155, 140/160, 135/165 and 100/200. Thus, all treatments share the same expected value (150) but differ in the size of the range of possible threshold values. Note that, for the above parameter values, $\phi = 15.8$. Our model thus predicts that players should avoid catastrophe in the 150 and 145/155 treatments but not in the 140/160, 135/165 and 100/200 treatments.

In total, 500 students participated in the computerized experiment, 100 per treatment (10 groups × 10 players per group). The games were played in stages. In the first stage, every participant proposed a contribution target for their group and pledged an amount they would contribute individually. It was common knowledge that these declarations were non-binding but would be communicated to the group. After these declarations were revealed, the participants chose their actual contributions in the second stage. All of the participants were then informed about everyone's decisions and asked to complete a short questionnaire, giving a picture of their reasoning and emotions during the game. Finally, a computerized 'spinning wheel' was activated to determine the actual value for the threshold.

The experimental results strongly support the hypotheses arising from our theoretical model (Fig. 1). In the 150 treatment, catastrophe is avoided eight out of ten times. In 145/155, catastrophe is avoided four out of ten times with probability 100% and in the other six cases with probability of at least 30%. The difference between 150 and 145/155 is statistically insignificant (Mann–Whitney–Wilcoxon test, $n = 20$, $p = 0.23$). In contrast, in 140/160, catastrophe is never avoided, despite the closeness of this uncertainty range to 145/155. Catastrophe also occurs every time in the 135/165 treatment. In 100/200, one group out of ten managed to reduce the probability of catastrophe seven per cent. Again, the differences among 140/160, 135/165 and 100/200 are insignificant ($n = 20$, $p > 0.05$ each). However, the differences between 150 and 145/155 on the one hand and 140/160, 135/165, and 100/200 on the other hand are all highly significant ($n = 20$, $p < 0.01$ each). Consistent with the theory, there is a qualitative change in behaviour either side of the dividing line.

As predicted, contributions in the treatments 150 and 145/155 do not differ significantly from the full cooperative levels of 150 and 155, respectively (see Table 1 and Fig. 2; t-test, $n = 10$, $p = 0.72$ for 150 and $p = 0.25$ for 145/155). Moreover, the average contribution in 145/155 is higher than in 150, although the difference lacks statistical significance (Mann–Whitney–Wilcoxon test, $n = 20$, $p = 0.85$).

Consistent with our theoretical model, contributions in the treatments 140/160, 135/165 and 100/200 are significantly different from their full cooperative levels—160, 165 and 200, respectively (t-test, $n = 10$, $p = 0.00$ each). However, contributions exceed the predicted value of zero (one-sided t-test, $n = 10$, $p = 0.00$ each). We should not be surprised. To the right of the dividing line, the players face a prisoners' dilemma, and it is a common finding in the experimental literature that groups playing this game contribute somewhere between the full cooperative and non-cooperative levels[22] (Supplementary Information Literature). In our experiment, the tendency to contribute above the non-cooperative level is especially strong because contributions from Account A are very cheap.

To the left of the dividing line, average proposals and pledges closely track their full cooperative levels; to the right, they fall relative to full cooperation as the uncertainty range increases (Table 1 and Fig. 2). Our ex post questionnaire (Supplementary Empirical Analysis) reveals that, to the left of the dividing line, proposals are mainly chosen to maximize the collective payoff of the group, pledges to signal intended contributions. To the right of the dividing line, in contrast, proposals and pledges are mainly chosen to stimulate contributions by others.

Figure 3 reveals the effect of uncertainty on individual pledges and contributions. To the left of the dividing line, contributions are tightly bunched near the full cooperative levels. To the right, contributions move progressively closer to zero and 10, the cheap-chips level, as the range of uncertainty increases. When the uncertainty range reaches 100/200, the full cooperative level has lost its attraction, with most players contributing zero or 10, and with only 2 players contributing 20. To the left of the dividing

Table 1 | Summary statistics by treatment.

Treatment	Proposal		Pledge		Contribution		Group contribution	
	Mean (s.d.)	Mode (%)	Mean (s.d.)	Mode (%)	Mean (s.d.)	Mode (%)	Mean (s.d.)	Min/max
150	151.9 (1.57)	150 (83%)	14.7 (0.51)	15 (74%)	15.1 (0.77)	15 (56%)	150.9 (7.69)	136/159
145/155	158.0 (1.40)	160 (48%)	15.1 (0.62)	16 (53%)	15.4 (0.38)	16 (45%)	153.5 (3.84)	148/160
140/160	161.0 (2.64)	160 (69%)	15.1 (0.83)	16 (64%)	11.7 (1.69)	16 (33%)	117.4 (16.85)	80/139
135/165	163.3 (8.75)	170 (41%)	15.4 (1.10)	17 (36%)	11.3 (1.98)	10 (33%)	112.9 (19.84)	68/130
100/200	166.3 (9.85)	200 (29%)	15.8 (1.69)	20 (32%)	7.7 (1.67)	10 (36%)	77.2 (16.67)	55/107

To the left of the dividing line for threshold uncertainty (top two rows), actual contributions closely follow the proposals and pledges. To the right of the dividing line (bottom three rows), contributions fall short of the proposals and pledges. Here, mean and especially modal proposals and pledges increase with the uncertainty range (that is, with the full cooperative contribution level), whereas mean and modal contributions decrease. The treatments 150 and 100/200 are taken from ref. 21.

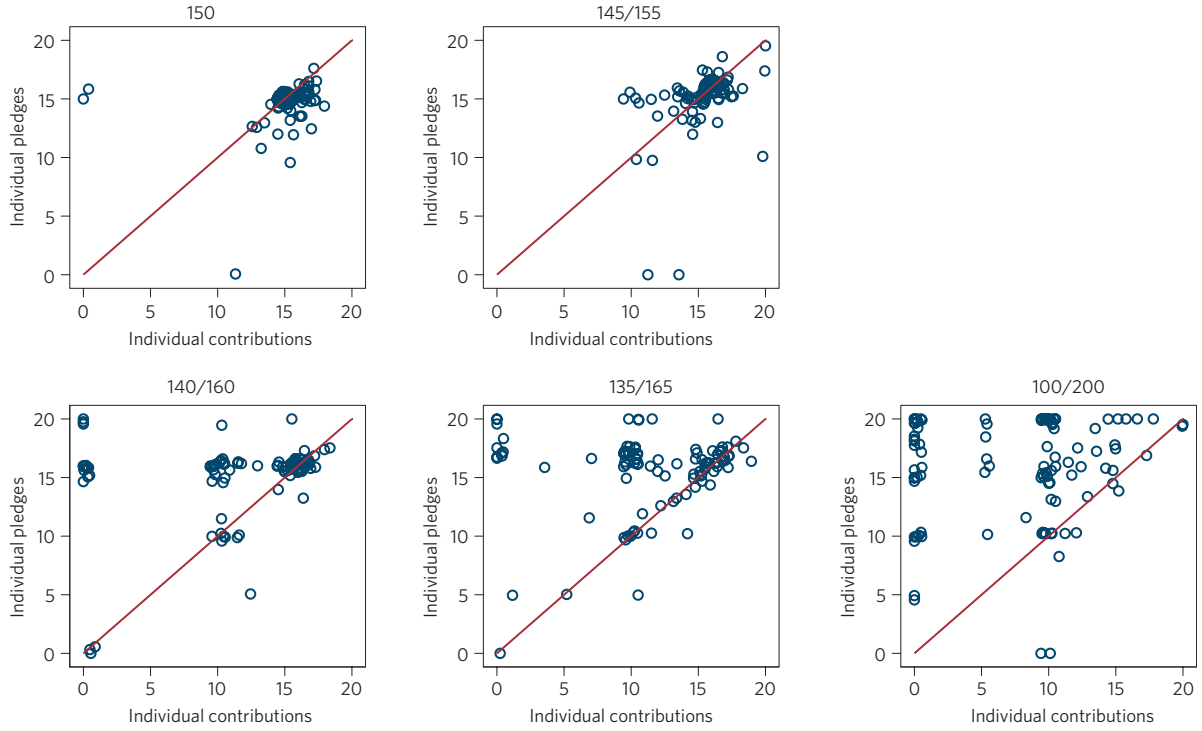

Figure 3 | Individual pledges and contributions by treatment. To the left of the dividing line, pledges and contributions are tightly bunched, more so in 150 than in 145/155. To the right of the dividing line, in 140/160, 135/165 and 100/200, values vary widely, with contributions increasingly falling short of pledges with higher uncertainty. A series of Spearman's correlation tests gives: $n = 100$, $\rho = 0.38$, $p = 0.00$ in 150; $\rho = 0.59$, $p = 0.00$ in 145/155; $\rho = 0.33$, $p = 0.00$ in 140/160; $\rho = -0.01$, $p = 0.93$ in 135/165; $\rho = 0.10$, $p = 0.34$ in 100/200. A small noise (2%) has been inserted to make all data points visible.

line, the players had incentives to be trustworthy and trusting, and most players contributed at least as much as they pledged (98% in 150 and 80% in 145/155). To the right of the dividing line, the incentives were different, and far fewer players contributed the amounts they pledged (55% in 140/160, 46% in 135/165, and only 18% in 100/200).

The contribution of poker chips in our experiment is best thought of as a metaphor for the things countries need to do to prevent 'dangerous' climate change. Ref. 23, for example, identified an atmospheric CO_2 concentration level of 350 ppmv as 'safe', on the basis of palaeoclimatic evidence suggesting that the polar ice sheets 'tipped' previously somewhere between 350 and 550 ppmv. Assuming that these values represent the range of the distribution for a critical threshold, our model can be interpreted, roughly, as suggesting that countries would do no better collectively than to limit concentrations to 350 ppmv, provided our first condition $(X \geq (c^{B} - bN)(\overline{Q}_{max} - Nq^{A}_{max})/N)$ held, but that they could be expected to behave so as to allow concentrations to top 550 ppmv unless our second condition $(X \geq (c^{B} - b)(\overline{Q}_{max} - \overline{Q}_{min}))$ also held. To give our analysis empirical real-world relevance, more sophisticated versions of these relations could be developed and estimated, but this will require further research[24].

Improved climate models[25–27] and early warning signals are valuable, not least as an aid to adaptation. By reducing threshold

uncertainty, they may also stimulate emergency measures to limit emissions. However, early warning signals may fail completely[8,28]; or, being prone to false positives and false negatives[9], may reduce uncertainty by too little to prevent a critical threshold from being breached. Making matters worse, early warning signals arrive late. Even if early warning changed the incentives to act, the ability to act may be severely circumscribed, leaving few immediate options besides adaptation and geoengineering.

The impact of crossing a tipping point can be interpreted as a punishment imposed by nature. When uncertainty about a tipping point is small, the fear of crossing it serves as an effective deterrent. When threshold uncertainty is large, however, this punishment fails as a deterrent, and strategic enforcement mechanisms are needed to deter players from straying into the climate 'danger zone'[29,30]. A good example of strategic enforcement is the use of trade restrictions in the Montreal Protocol, which has successfully protected the ozone layer. Similar mechanisms, our research implies, are needed to avoid 'dangerous' climate change.

Methods

The two key conditions of our theoretical model result from two different optimization exercises. If countries cooperate fully, they will choose their individual abatement levels so as to maximize their expected aggregate payoff, $E(\Pi) = bQN - \sum_i c^A q_i^A - \sum_i c^B q_i^B - XN[1-(Q-\overline{Q}_{min})/(\overline{Q}_{max}-\overline{Q}_{min})]$ for $Q \in [\overline{Q}_{min}, \overline{Q}_{max}]$. Under our assumptions, it will pay all countries collectively to abate $\overline{Q}_{max}$ in the aggregate, just, so long as $XN \geq (c^B - bN)(\overline{Q}_{max} - Nq_{max}^A)$ (Supplementary Methodological Details). We assume that this condition is always satisfied. If countries choose their abatement levels independently, every country i will maximize its own expected payoff, $E(\pi_i) = bQ - c^A q_i^A - c^B q_i^B - X[1-(Q-\overline{Q}_{min})/(\overline{Q}_{max}-\overline{Q}_{min})]$ for $Q \in [\overline{Q}_{min}, \overline{Q}_{max}]$, taking the abatement levels of other countries as given. Zero abatement is always a Nash equilibrium, but per-country abatement in the amount $\overline{Q}_{max}/N$ is also a Nash equilibrium when $X \geq (c^B - b)(\overline{Q}_{max} - \overline{Q}_{min})$ (Supplementary Methodological Details). This last condition is satisfied in two of our experimental treatments (150 and 145/155); it is not satisfied for the three remaining treatments (140/160, 135/165 and 100/200).

The experimental sessions were conducted in a computer laboratory at the University of Magdeburg, using students recruited from the general student population. In each session, subjects were seated randomly at linked computers. A set of written instructions including several numerical examples and control questions was handed out. 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. Note that there is no significant correlation between the average contributions made in the practice rounds and the contributions made in the real round (in every case, the p values were insignificant at the 10% level). To ensure anonymity, each member of a group was identified by a different letter (A–J). At the end of each session, after the actual threshold value was determined by the 'spinning wheel,' students were paid their earnings in cash (for more details see the Supplementary Methodological Details).

Received 23 July 2013; accepted 30 October 2013; published online 8 December 2013

References

1. Alley, R. B. *et al.* Abrupt climate change. *Science* **299,** 2005–2010 (2003).
2. Lenton, T. M. *et al.* Tipping elements in the earth's climate system. *Proc. Natl Acad. Sci. USA* **105,** 1786–1793 (2008).
3. Kriegler, E., Hall, J. W., Held, H., Dawson, R. & Schellnhuber, H. J. Imprecise probability assessment of tipping points in the climate system. *Proc. Natl Acad. Sci. USA* **106,** 5041–5046 (2009).
4. Dakos, V. *et al.* Slowing down as an early warning signal for abrupt climate change. *Proc. Natl Acad. Sci. USA* **105,** 14308–14312 (2008).
5. Biggs, R., Carpenter, S. R. & Brock, W. A. Turning back from the brink: Detecting an impending regime shift in time to avert it. *Proc. Natl Acad. Sci. USA* **106,** 826–831 (2009).
6. Scheffer, M. *Critical Transitions in Nature and Society* (Princeton Univ. Press, 2009).
7. Scheffer, M. *et al.* Early-warning signals for critical transitions. *Nature* **461,** 53–59 (2009).
8. Ditlevsen, P. D. & Johnsen, S. J. Tipping points: Early warning and wishful thinking. *Geophys. Res. Lett.* **37,** L19703 (2010).
9. Lenton, T. M. Early warning of climate tipping points. *Nature Clim. Change* **1,** 201–209 (2011).
10. Scheffer, M. *et al.* Anticipating critical transitions. *Science* **338,** 344–348 (2012).
11. Lenton, T. M., Livina, V. N., Dakos, V., van Nes, E. H. & Scheffer, M. Early warning of climate tipping points from critical slowing down: Comparing methods to improve robustness. *Phil. Trans. R. Soc. A.* **370,** 1185–1204 (2012).
12. Carpenter, S. R. *et al.* Early warnings of regime shifts: A whole-ecosystem experiment. *Science* **332,** 1079–1082 (2011).
13. Wang, R. *et al.* Flickering gives early warning signals of a critical transition to a eutrophic lake state. *Nature* **492,** 419–422 (2012).
14. Scheffer, M., Carpenter, S., Foley, J. A., Folke, C. & Walker, B. Catastrophic shifts in ecosystems. *Nature* **413,** 591–596 (2001).
15. Dai, L., Vorselen, D., Korolev, K. S. & Gore, J. Generic indicators for loss of resilience before a tipping point leading to population collapse. *Science* **336,** 1175–1177 (2012).
16. May, R. M., Levin, S. A. & Sugihara, G. Complex systems: Ecology for bankers. *Nature* **451,** 893–895 (2008).
17. Lackner, K. S. *et al.* The urgency of the development of CO_2 capture from ambient air. *Proc. Natl Acad. Sci. USA* **109,** 13156–13162 (2012).
18. Allen, M. R. *et al.* Warming caused by cumulative carbon emissions towards the trillionth tonne. *Nature* **458,** 1163–1166 (2009).
19. Zickfeld, K., Eby, M., Matthews, H. D. & Weaver, A. J. Setting cumulative emissions targets to reduce the risk of dangerous climate change. *Proc. Natl Acad. Sci. USA* **106,** 16129–16134 (2009).
20. Barrett, S. Climate treaties and approaching catastrophes. *J. Environ. Econ. Manage.* **66,** 235–250 (2013).
21. Barrett, S. & Dannenberg, A. Climate negotiations under scientific uncertainty. *Proc. Natl Acad. Sci. USA* **109,** 17372–17376 (2012).
22. Ledyard, J. O. in *Handbook of Experimental Economics* (eds Kagel, J. H. & Roth, A. E.) 111–194 (Princeton Univ. Press, 1995).
23. Rockström, J. *et al.* A safe operating safe for humanity. *Nature* **461,** 472–475 (2009).
24. Lenton, T.M. & Ciscar, J-C. Integrating tipping points into climate impact assessments. *Climatic Change* **117,** 585–597.
25. Robinson, A., Calov, R. & Ganopolski, A. Multistability and critical thresholds of the Greenland ice sheet. *Nature Clim. Change* **2,** 429–432 (2012).
26. Hawkins, E. *et al.* Bistability of the Atlantic overturning circulation in a global climate model and links to ocean freshwater transport. *Geophys. Res. Lett.* **38,** L10605 (2011).
27. Drijfhout, S.S., Weber, S.L. & van der Swaluw, E. The stability of the MOC as diagnosed from model projections for pre-industrial, present and future climates. *Clim. Dynam.* **37,** 1575–1586 (2010).
28. Hastings, A. & Wysham, D. B. Regime shifts in ecological systems can occur with no warning. *Ecol. Lett.* **13,** 464–472 (2010).
29. Ostrom, E. *Governing the Commons: The Evolution of Institutions for Collective Action* (Cambridge Univ. Press, 1990).
30. Barrett, S. *Environment and Statecraft: The Strategy of Environmental Treaty-Making* (Oxford Univ. Press, 2003).

Acknowledgements

We thank J. Rising for programming our 'spinning wheel', the MaXLab team at Magdeburg University for use of their laboratory and the Princeton Institute for International and Regional Studies research community on Communicating Uncertainty: Science, Institutions, and Ethics in the Politics of Global Climate Change for financially supporting our experiments.

Author contributions

S.B. and A.D. contributed equally to this work. They both designed and performed the research and analysed the data and wrote the paper.

Additional information

Supplementary information is available in the online version of the paper. Reprints and permissions information is available online at www.nature.com/reprints. Correspondence and requests for materials should be addressed to S.B.

Competing financial interests

The authors declare no competing financial interests.