SUPPLEMENTARY MATERIALS:

Threshold uncertainty, early-warning signals, and the prevention of dangerous climate change

Mark J. Hurlstone[1*] and Ben R. Newell[2]

[1]School of Psychological Science, University of Western Australia, Perth, Australia
[2]School of Psychology, UNSW, Sydney, Australia
[*]mark.hurlstone@uwa.edu.au

Hurlstone and Newell[1] experimentally examined whether early-warning signals of approaching climate thresholds could overturn the impediment of threshold uncertainty on cooperation in a catastrophe avoidance threshold public goods game. They showed that large initial—and subsequently unabated—threshold uncertainty undermines cooperation, consistent with earlier studies,[2–4] but additionally that a marked subsequent reduction in threshold uncertainty—mimicking an early-warning signal—does little to improve the prospects of cooperation. The findings suggest that early-warning signals indicating that a critical climate threshold is approaching are unlikely to offer the leverage necessary to motivate countries to take the necessary action to avert catastrophe. This supplementary document reports additional details about the study including an extended literature review; the instructions given to participants; and ancillary statistical analyses. Note that this document is not meant to be self-explanatory—please consult Hurlstone and Newell[1] for further information.

1 Supplementary Literature

A literature has recently developed that uses cooperation games to simulate aspects of the international negotiations on climate change in the experimental laboratory. The behavioural insights gleaned from this literature have been obtained using a catastrophe avoidance game developed by Milinski et al.[5] known as known as the "collective-risk social dilemma" (CRSD). The game involves groups of six players. Each is given an operating fund of €40 and must decide whether to contribute €0, €2, or €4 in each of 10 rounds to a climate account without communicating. At the end of each round, the contributions of each group member are made public. If at least €120 has been contributed by the end of the game then catastrophic climate change is averted with certainty and players get to keep the leftovers of their operating fund. However, if the group fails to reach the threshold then catastrophic climate change occurs with a pre-specified probability (e.g., 90 %) that what remains of each player's operating fund will be lost. The €120 target can be construed as a temperature threshold—such as the 2°C goal—whilst the player contributions are a metaphor for the level of investment of countries in emission reductions. The CRSD is a coordination game—where players must coordinate strategies for their mutual benefit—with two symmetrical pure strategy Nash equilibria. One is a "dangerous" equilibrium where each player contributes €0, whereas the other is a "safe" equilibrium where each player contributes €20 (there are also several "safe" asymmetric pure strategy Nash equilibria where different players contribute different amounts). Coordination games usually possess a "focal point"[6] with salient characteristics that facilitates coordination. In the CRSD, fear of catastrophe makes the €120 contribution level salient, rendering it a natural focal point.

Using this game—and basic variants of it[2,3,7]—the impact of a number of variables on the climate negotiations has been examined. The perception of risk—the probability of catastrophe if the threshold is crossed—has been identified as a major driver of cooperation—when the perception of risk is low or moderate, cooperation is difficult to achieve, but when the perception of risk is high, groups can easily cooperate to keep on the safe side of the threshold.[5] Inequalities in historical responsibility,[8] wealth,[9,10] and risk exposure[10] have weak to moderate negative effects, whereas intergenerational discounting is a serious handicap to cooperation.[11] The use of naming and shaming of defectors in a bid to facilitate cooperation—in a process akin to the Paris pledge-and-review enforcement mechanism—has no reliable effect on group contributions.[7] Of specific relevance to our own research is the effects of uncertainties about the location of the threshold and the damages of crossing it, which have been investigated in a series of theoretical and empirical studies by Barrett and Dannenberg.[2–4,7,12] These authors have shown theoretically[12] and experimentally[13] that the existence of a dangerous climate threshold facilitates cooperation, relative to a scenario based on gradual climate change alone. However, uncertainty surrounding the threshold causes cooperation to collapse,[2–4] whereas uncertainty about the damages has no effect on behavior provided that the damages of crossing the threshold exceed the costs of avoiding it.[2]

In the first of these studies, Barrett and Dannenberg[2] jointly examined the influence of uncertainty about the threshold for catastrophe and the damages of crossing it on cooperation. In their variant of the CRSD, groups consist of ten players who are each allocated €31, which is divided into an operating fund of €11 and an endowment of €20. The operating fund can be used to invest in "weak" or "strong" emission abatement by purchasing poker chips (max = 10 of each type) at a cost of €0.10 or €1.00, respectively. The game is played over a single round which is divided into two stages—a communication stage, where each player submits a proposal regarding the contribution target for the group and pledges an amount they will contribute individually, followed by a contribution stage where each player chooses their actual contributions. The goal is to reach an investment threshold T, otherwise a cost C is deducted from each player's endowment with certainty (100 % risk of catastrophe).

Barrett and Dannenberg's experiment contained four treatments. In the certainty treatment, the dangerous threshold and the impact of crossing it were both known with certainty (T = 150 and C = €15). In the damage uncertainty treatment, the dangerous threshold was known with certainty, but the damage of crossing it was not (T = 150 and C was uniformly distributed between €10–€20). In the threshold uncertainty treatment, the damage of crossing the dangerous threshold was known with certainty, but the location of the threshold was not (T was uniformly distributed between 100–200, and C = €15). Finally, in the damage-and-threshold uncertainty treatment, both the dangerous threshold and the damage of crossing it were uncertain (T was uniformly distributed between 100–200, and C was uniformly distributed between €10–€20). When the threshold and damages were known with certainty, 80 % of groups averted catastrophe and this figure rose to 100 % in the damage uncertainty treatment, but only 10 % and 30 % of groups in the threshold uncertainty and damage-and-threshold uncertainty treatments, respectively, averted catastrophe (Fig. S1a). These results demonstrate that uncertainty surrounding the threshold for dangerous climate change is a major handicap to cooperation, whereas uncertainty about the damages is inconsequential.

These results were obtained using a one-shot interaction game, in which all decisions were made during a single round of play. However, Dannenberg et al.[4] have generalised these results to a repeated interaction scenario using the original Milinski CRSD game. One augmentation of this game introduced by the authors was to incorporate a communication component akin to that used in the Barrett and Dannenberg study.[2] This was achieved by allowing players to submit nonbinding pledges regarding how much they intended to contribute in the catastrophe avoidance game over the total of ten rounds at the start of round 1, and again on round 6. Dannenberg et al. compared three different treatments: a certainty treatment in which the threshold was €120, an uncertainty treatment in which the threshold was a uniform random variable between €0–€240, and an ambiguity treatment in which the threshold was also between €0–€240 but the underlying probability distribution was unknown. Cooperation across the three treatments was indexed by comparing the number of groups that successfully reached various different hypothetical thresholds (Fig. S1b). At hypothetical threshold values of €20, €40, and €60, group success rates were at, or close to, ceiling and did not vary appreciably across the three treatments, but at a hypothetical threshold of €80, ambiguity began to exert a modest negative effect on group success rates. At a hypothetical threshold of €100, both uncertainty and ambiguity exerted a modest negative effect on group success rates, whilst at a hypothetical threshold of €120, both uncertainty and ambiguity exerted a strong negative effect on group success rates.

Why does threshold uncertainty undermine cooperation? Using a game theoretic model, Barrett[12] has shown that threshold uncertainty changes the nature of the game being played. Specifically, along the threshold uncertainty range there exists a hypothetical dividing line which distinguishes between two types of games. To the left of the dividing line, when the threshold is certain or the uncertainty about the threshold is only very narrow, the game being played is a coordination game, whereas to the right of the dividing line—beyond this narrow level of uncertainty—the game being played is a prisoners' dilemma. The key difference between a coordination game and a prisoners' dilemma is that the latter game is characterised by a single noncooperative pure strategy Nash equilibrium towards which players will gravitate. Accordingly, in the prisoners' dilemma, mutual defection—rather than mutual cooperation—becomes the focal and dominant strategy. Thus, the prediction from game theory is that there is some degree of tolerance in the level of uncertainty surrounding the threshold. It may not, therefore, be necessary to eliminate threshold uncertainty entirely to improve the prospects of cooperation, just enough to transform the prisoners' dilemma game back into a coordination game under which cooperation is more readily attainable.

Barrett and Dannenberg[3] conducted an experiment to test the predictions of this game theoretic account. Using the one-shot interaction game employed in their earlier study,[2] they varied the size of the window of uncertainty surrounding the threshold. In the certainty treatment, T = 150, whereas in four threshold uncertainty treatments T was uniformly distributed between either: (1) 145–155, (2) 140–160, (3) 135–165, or (4) 100–200 (C = €15 in all treatments)[1]. Based on the theory of Barrett,[12] it was predicted that to the left of the dividing line, most groups would avert catastrophe in treatment 150, with somewhat fewer groups doing so in treatment 145–155. By contrast, to the right of the dividing line, it was predicted that most groups

[1] In practice, the data for treatments 150 and 100–200 were taken from the certainty and uncertainty treatments of Barrett and Dannenberg.[2]

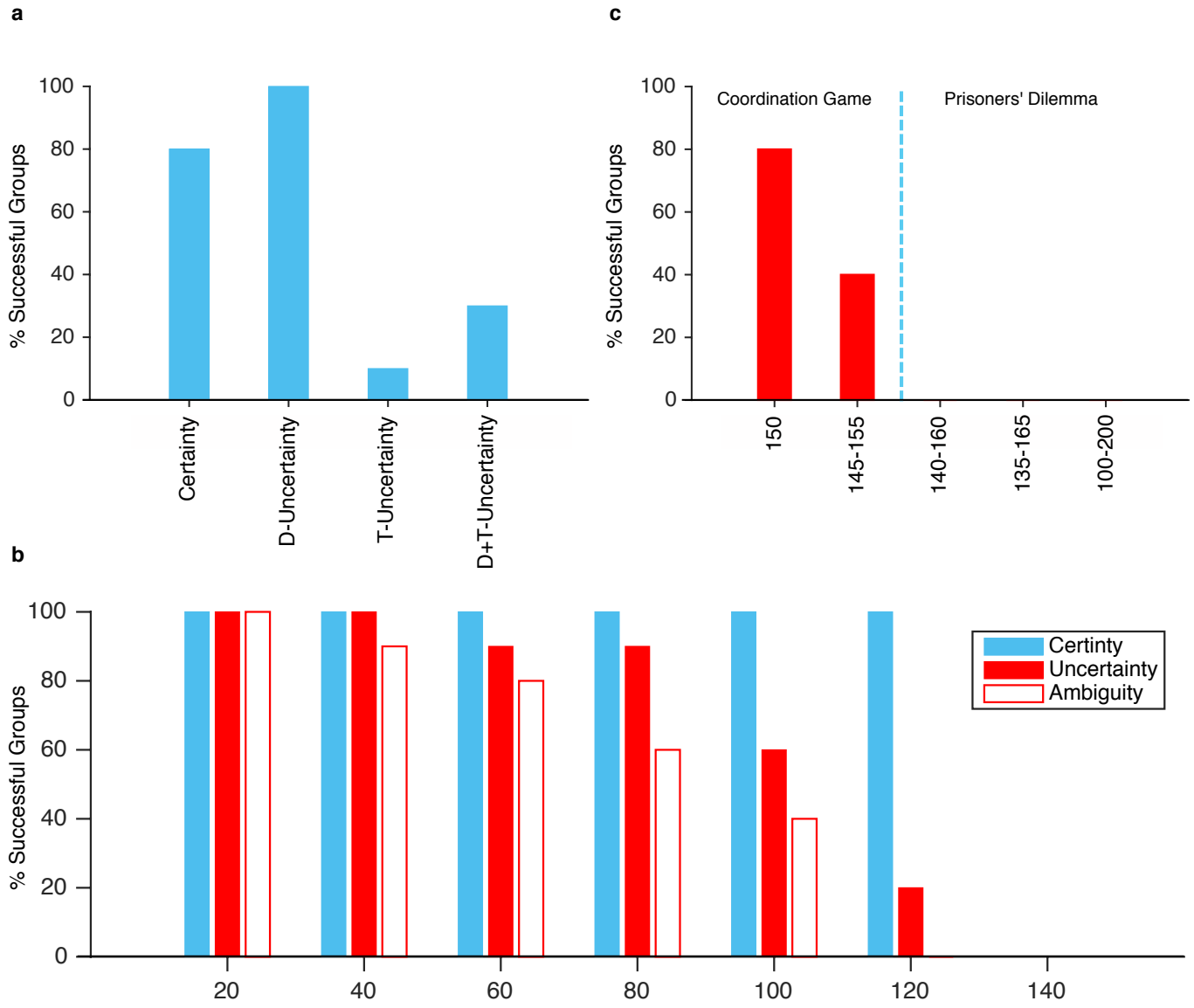

Figure S1 | Percentage of groups reaching the threshold as a function of treatment in three different studies. a, Barrett and Dannenberg (2012),[2] **b,** Dannenberg et al. (2015),[4] and **c,** Barrett and Dannenberg (2013).[3]

would fail to avert catastrophe. The results shown in Fig. S1c confirm these qualitative predictions—to the left of the dividing line, 80 % of groups avoided catastrophe in treatment 150, and this figure dropped to 40 % in treatment 145–155. However, to the right of the dividing line, all groups failed to avert catastrophe in the three remaining uncertainty treatments. These results suggest that efforts to reduce uncertainty about the proximity of a dangerous climate threshold might catalyse action to avoid it.

Recently it has been suggested that generic early-warning signals may exist that signify we are approaching a climate threshold.[14–18] Based on their findings, Barrett and Dannenberg[3] noted that the failure of collective action when threshold uncertainty is large might be surmounted if an early-warning signal could be detected that reduced threshold uncertainty sufficiently. Based on their results, such a signal would need to reduce threshold uncertainty to within less than 10 % of the true threshold, although it is questionable whether an early-warning signal could provide such a high level of precision.[19] Moreover, the one-shot interaction game employed by Barrett and Dannenberg[3] does not provide a direct empirical test of this early-warning signal hypothesis. In their game, proposals, pledges, and group contributions unfold over a single round of play, and group members always face the same level of threshold uncertainty (threshold uncertainty is varied between treatments, but does not change over time within the same treatment). There is no opportunity for players to have their beliefs about what other players will contribute to be confirmed or disconfirmed, and there is no opportunity to negotiate a new collective target given updated information about the

proximity of the group to the threshold. A direct test of the early-warning signal hypothesis requires a repeated interaction game in which groups play the first half of the game under large threshold uncertainty, before mid-game receiving an early-warning signal that the uncertainty about the threshold has been reduced, before then completing the second half of the game. Under this repeated interaction scenario, we might expect an early-warning signal—even one that reduces uncertainty to within 10 % of the true threshold value—to be far less effective at catalysing cooperation. This is because the large threshold uncertainty faced by groups initially might cause cooperation to collapse to a point from which recovery is no longer possible given the remaining time available, irrespective of the degree of precision of the early-warning signal. Using an augmented version of Milinski's[5] multiple round CRSD that permitted communication between players, this is the gap in the literature that our study attempts to fill.

2 Supplementary Experimental Instructions (adapted from[4])

Treatment T_1

Welcome to our experiment!

1. General Information

In our experiment, you can earn money. How much you earn depends on the gameplay, or more precisely on the decisions you and your fellow players make. Regardless of the gameplay, you will receive $10 for your participation. For a successful experiment, it is necessary that you do not talk to other participants or do not communicate in any other way. Now please read the following rules of the game carefully. If you have any questions, please raise your hand.

2. Game Rules

There are six players in the game, meaning you and five other players. Each player is faced with the same decision problem. In the beginning of the experiment, you receive a starting capital of $40, which is credited to your personal account. During the experiment, you can use the money in your account or let it be. In the end, your current account balance is paid to you in cash. Your decisions are anonymous. For the purpose of anonymity, you will be allocated a pseudonym which will be used for the whole duration of the game. The pseudonyms are chosen from the names of moons in the Solar System (Ananke, Telesto, Despina, Japetus, Kallisto or Metis). Once the game begins you will be able to see your pseudonym in the lower left corner of your display.

The experiment has exactly ten rounds. In each round, you can invest your money in order to try and prevent damage. The damage will have a considerable negative financial impact on all players. In each round of the game, all six players are asked the following question at the same time:

"How much do you want to invest to prevent damage?"

You can answer with $0, $2, or $4. After each player has made her or his decision, the six decisions are displayed at the same time. After that, all money paid by the players is assigned to a special account for damage prevention.

At the end of the game (after exactly ten rounds), the computer calculates the total investments made by all players of the group. If the total investments are equal to or greater than a threshold amount, the damage is prevented and each player is paid the money remaining in her or his account, meaning the $40 starting capital minus the money the player has invested in preventing damage over the course of the game. However, if the total investments are lower than the threshold amount, the damage occurs: All players lose 90% of the remaining money in their personal accounts. The threshold amount to be reached in order to prevent damage is $120.

At the end of the game all players together must have invested at least $120 to prevent the damage. If a single player has invested, say, a total of $10 in damage prevention after ten rounds, he or she has a credit of $30 on his or her personal account. If the group of players as a whole has invested $120 or more in damage prevention, the damage will not occur and this player will receive $30 from the game. However, if the group has invested less than $120, the damage will occur and the player will receive $3 (10% of $30) from the game.

Please note the following feature of the game: Before the players decide how much they want to invest into preventing damage, they make two non-binding announcements. First, each player makes a proposal for how much the group as a whole should invest into preventing damage over the total of ten rounds. Second, each player makes a pledge for how much money they intend to invest in total over the ten rounds into preventing damage. After these two non-binding announcements, the proposals and pledges made by all players (and an average and total value from all proposals and pledges, respectively) will be shown on the monitor. At the end of round 5, all players can make a new non-binding proposal for the total investments to be made by the group over the ten rounds, and a new non-binding pledge for how much money they intend to invest in total over the ten rounds.

3. An Example

Here, you can see an example of the decisions made by the six players in one round (round 3). Please direct your attention first to the far right of the graphic.

The fourth column shows the investments made in the current round (round 3). The players Ananke and Kallisto have invested $2 each, the players Telesto and Japetus have invested $4 each and Despina and Metis have not made any investments. In total, $12 were invested in this round. The third column shows the cumulative investments made by each player from the first to the current round (rounds 1–3). The players Ananke and Telesto have each invested $6 in the first three rounds. Despina, Kallisto and

Proposals Rounds 1-10		Pledges Rounds 1-10		Investments Rounds 1-3		Investments Round 3	
Ananke	100	Ananke	10	Ananke	6	Ananke	2
Telesto	80	Telesto	12	Telesto	6	Telesto	4
Despina	120	Despina	20	Despina	4	Despina	0
Japetus	100	Japetus	12	Japetus	10	Japetus	4
Kallisto	110	Kallisto	22	Kallisto	4	Kallisto	2
Metis	140	Metis	24	Metis	4	Metis	0
Average	108	Total	100	Total	34	Total Round 3	12

Metis have each invested $4 and Japetus has invested $10 in the first three rounds. In total, $34 were invested in the first three rounds.

The first column shows the proposals made by each player regarding how much the group as a whole should invest into preventing damage over the ten rounds in total. For example, Metis suggests that the group should invest $140. The average of all proposals is $108. The second column shows the pledges made by each player regarding how much they will personally invest in the damage prevention account over the ten rounds in total. For example, over the ten rounds Kallisto has pledged to personally invest $22 in total. The total of all pledges is $100. In the game, you will see this information after each round.

4. Control Questions

1. How much does a player have to invest on average over the course of ten rounds, if the group was to invest $120 in total?

☐ $10 ☐ $12 ☐ $20 ☐ $30 ☐ $60

2. Assume the group has invested the threshold amount to prevent damage, and that you have invested $16 in total. How much cash do you get at the end of the game (excluding the $10 participation fee)?

I get $ _____.

3. Take a look at the table in part 3 of the instructions.

(a) How much did Ananke and Kallisto propose the group should invest in damage prevention over the ten rounds?

Ananke proposed $_____. Kallisto proposed $_____.

(b) How much did Japetus and Metis pledge to invest in damage prevention over the ten rounds?

Japetus pledged $_____. Metis pledged $_____.

(c) How much money do Despina and Japetus have in their personal accounts after round 3?

Despina has $_____ in her account. Japetus has $_____ in his account.

4. True or false? At the start of the game, and once again at the end of round 5, each player makes: (I) a non-binding proposal of how much the group should collectively invest in damage prevention over the ten rounds, and (II) a non-binding pledge of how much they will personally invest in damage prevention over the ten rounds.

☐ True ☐ False

5. Assume you invested a total of $20 over the ten rounds and the threshold amount was not reached by your group. How much cash do you get at the end of the game (excluding the $10 participation fee)?

☐ $0 ☐ $2 ☐ $4 ☐ $10 ☐ $20

6. Assume that the group has invested a total of $100 over the ten rounds. Does the damage occur in this case? (please tick the correct answer).

☐ Yes ☐ No

7. Assume that the group has invested a total of $125 over the ten rounds. Does the damage occur in this case? (please tick the correct answer).

☐ Yes ☐ No

Please raise your hand after you have answered all control questions. We will come to you and check the answers. The game will begin after we have checked the answers of all players and answered any questions you may have. Good luck!

Treatments T_2, T_3, and T_4

Welcome to our experiment!

1. General Information

In our experiment, you can earn money. How much you earn depends on the gameplay, or more precisely on the decisions you and your fellow players make. Regardless of the gameplay, you will receive $10 for your participation. For a successful experiment, it is necessary that you do not talk to other participants or do not communicate in any other way. Now please read the following rules of the game carefully. If you have any questions, please raise your hand.

2. Game Rules

There are six players in the game, meaning you and five other players. Each player is faced with the same decision problem. In the beginning of the experiment, you receive a starting capital of $40, which is credited to your personal account. During the experiment, you can use the money in your account or let it be. In the end, your current account balance is paid to you in cash. Your decisions are anonymous. For the purpose of anonymity, you will be allocated a pseudonym which will be used for the whole duration of the game. The pseudonyms are chosen from the names of moons in the Solar System (Ananke, Telesto, Despina, Japetus, Kallisto or Metis). Once the game begins you will be able to see your pseudonym in the lower left corner of your display.

The experiment has exactly ten rounds. In each round, you can invest your money in order to try and prevent damage. The damage will have a considerable negative financial impact on all players. In each round of the game, all six players are asked the following question at the same time:

"How much do you want to invest to prevent damage?"

You can answer with $0, $2, or $4. After each player has made her or his decision, the six decisions are displayed at the same time. After that, all money paid by the players is assigned to a special account for damage prevention.

At the end of the game (after exactly ten rounds), the computer calculates the total investments made by all players of the group. If the total investments are equal to or greater than a threshold amount, the damage is prevented and each player is paid the money remaining in her or his account, meaning the $40 starting capital minus the money the player has invested in preventing damage over the course of the game. However, if the total investments are lower than the threshold amount, the damage occurs: All players lose 90% of the remaining money in their personal accounts. The threshold amount to be reached in order to prevent damage is some amount between $0 and $240, but you will not know the exact amount until the conclusion of the game. At the end of the experiment, the exact threshold amount will be drawn randomly by the computer. The draw is programmed so that each whole dollar amount between $0 and $240 has an equal probability of being selected.

Suppose at the end of the game that the randomly drawn threshold amount is $100. All players together must have invested at least $100 to prevent the damage. If a single player has invested, say, a total of $10 in damage prevention after ten rounds, he or she has a credit of $30 on his or her personal account. If the group of players as a whole has invested $100 or more in damage prevention, the damage will not occur and this player will receive $30 from the game. However, if the group has invested less than $100, the damage will occur and the player will receive $3 (10% of $30) from the game.

Please note the following feature of the game: Before the players decide how much they want to invest into preventing damage, they make two non-binding announcements. First, each player makes a proposal for how much the group as a whole should invest into preventing damage over the total of ten rounds. Second, each player makes a pledge for how much money they intend to invest in total over the ten rounds into preventing damage. After these two non-binding announcements, the proposals and pledges made by all players (and an average and total value from all proposals and pledges, respectively) will be shown on the monitor. At the end of round 5, all players can make a new non-binding proposal for the total investments to be made by the group over the ten rounds, and a new non-binding pledge for how much money they intend to invest in total over the ten rounds.

3. An Example

Here, you can see an example of the decisions made by the six players in one round (round 3). Please direct your attention first to the far right of the graphic.

The fourth column shows the investments made in the current round (round 3). The players Ananke and Kallisto have invested $2 each, the players Telesto and Japetus have invested $4 each and Despina and Metis have not made any investments. In total, $12 were invested in this round. The third column shows the cumulative investments made by each player from the first to the

Proposals Rounds 1-10		Pledges Rounds 1-10		Investments Rounds 1-3		Investments Round 3	
Ananke	100	Ananke	10	Ananke	6	Ananke	2
Telesto	80	Telesto	12	Telesto	6	Telesto	4
Despina	120	Despina	20	Despina	4	Despina	0
Japetus	100	Japetus	12	Japetus	10	Japetus	4
Kallisto	110	Kallisto	22	Kallisto	4	Kallisto	2
Metis	140	Metis	24	Metis	4	Metis	0
Average	108	Total	100	Total	34	Total Round 3	12

current round (rounds 1–3). The players Ananke and Telesto have each invested $6 in the first three rounds. Despina, Kallisto and Metis have each invested $4 and Japetus has invested $10 in the first three rounds. In total, $34 were invested in the first three rounds.

The first column shows the proposals made by each player regarding how much the group as a whole should invest into preventing damage over the ten rounds in total. For example, Metis suggests that the group should invest $140. The average of all proposals is $108. The second column shows the pledges made by each player regarding how much they will personally invest in the damage prevention account over the ten rounds in total. For example, over the ten rounds Kallisto has pledged to personally invest $22 in total. The total of all pledges is $100. In the game, you will see this information after each round.

4. Control Questions

1. How much does a player have to invest on average over the course of ten rounds, if the group was to invest $60 in total?

☐ $10 ☐ $12 ☐ $20 ☐ $30 ☐ $60

2. How much does a player have to invest on average over the course of ten rounds, if the group was to invest $180 in total?

☐ $10 ☐ $12 ☐ $20 ☐ $30 ☐ $60

3. Assume the group has invested the threshold amount to prevent damage, and that you have invested $16 in total. How much cash do you get at the end of the game (excluding the $10 participation fee)?

I get $ _____.

4. Take a look at the table in part 3 of the instructions.

(a) How much did Ananke and Kallisto propose the group should invest in damage prevention over the ten rounds?

Ananke proposed $_____. Kallisto proposed $_____.

(b) How much did Japetus and Metis pledge to invest in damage prevention over the ten rounds?

Japetus pledged $_____. Metis pledged $_____.

(c) How much money do Despina and Japetus have in their personal accounts after round 3?

Despina has $_____ in her account. Japetus has $_____ in his account.

5. True or false? At the start of the game, and once again at the end of round 5, each player makes: (I) a non-binding proposal of how much the group should collectively invest in damage prevention over the ten rounds, and (II) a non-binding pledge of how much they will personally invest in damage prevention over the ten rounds.

☐ True ☐ False

6. True or false? In the random draw to determine the threshold amount at the end of the game, each whole dollar amount between $0 and $240 has the same probability of being selected.

☐ True ☐ False

7. Assume you invested a total of $20 over the ten rounds and the threshold amount was not reached by your group. How much cash do you get at the end of the game (excluding the $10 participation fee)?

☐ $0 ☐ $2 ☐ $4 ☐ $10 ☐ $20

8. Assume that the group has invested a total of $100 over the ten rounds. The draw shows that the threshold amount to avoid damage is $160. Does the damage occur in this case? (please tick the correct answer).

☐ Yes ☐ No

9. Assume that the group has invested a total of $80 over the ten rounds. The draw shows that the threshold amount to avoid damage is $20. Does the damage occur in this case? (please tick the correct answer).

☐ Yes ☐ No

10. What is the probability of the threshold amount to prevent damage being greater than $60? _____.

11. What is the probability of the threshold amount to prevent damage being greater than $180? _____.

Please raise your hand after you have answered all control questions. We will come to you and check the answers. The game will begin after we have checked the answers of all players and answered any questions you may have. Good luck!

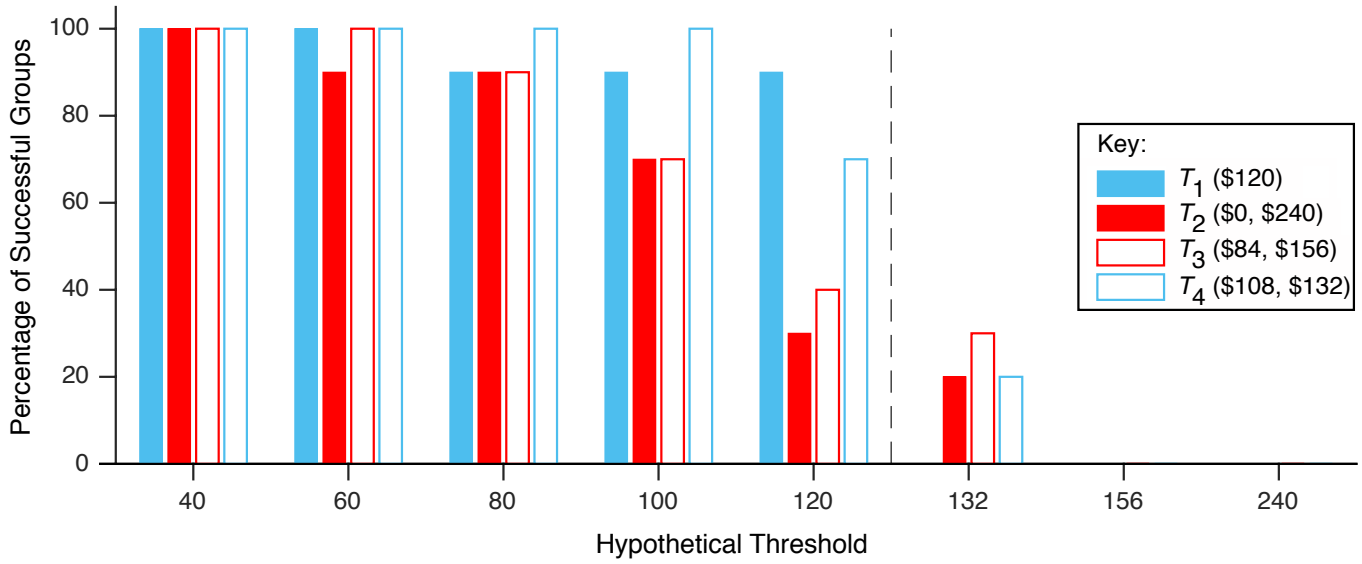

Figure S2 | Percentage of successful groups, as a function of treatment, for various hypothetical thresholds.

3 Supplementary Statistical Analyses

3.1 Success at reaching various hypothetical thresholds

The percentage of groups that would have averted catastrophe at various hypothetical thresholds is shown in Fig. S2. At threshold values of \$40, \$60, and \$80, most groups would have averted catastrophe, irrespective of treatment. At a threshold value of \$100, 90% of groups in T_1, 70% of groups in T_2 and T_3, and 100% of groups in T_4 would have averted catastrophe. At the focal[6] threshold value of \$120, 90% of groups in T_1, 30% of groups in T_2, 40% of groups in T_3, and 70% of groups in T_4 would have averted catastrophe. Success rates at this threshold were compared across treatments using contingency tables. Success rates were significantly higher in T_1 than in T_2 (Fisher exact, $P = 0.020$), whereas the difference between T_2 and T_3 (Fisher exact, $P = 1.000$) and between T_2 and T_4 (Fisher exact, $P = 0.179$) were both nonsignificant. Thus, at the \$120 threshold value, neither a 70% or 90% reduction in threshold uncertainty significantly increased the likelihood of averting catastrophe.

Fig. S2 shows group success rates at three additional hypothetical thresholds, namely \$132, \$156, and \$240. These correspond to the upper threshold limits that groups must have reached in the T_4, T_3, and T_2 treatments, respectively, to avert catastrophe with certainty. At \$132, only 20% of groups in T_2, 30% of groups in T_3, and 20% of groups in T_4 would have averted catastrophe. That more groups in T_4 did not reach the \$132 threshold is noteworthy, given that a fair-share contribution of \$22 per player would have ensured that catastrophe was averted with certainty. At \$156 and \$240, none of the groups would have averted catastrophe.

3.2 Contributions over rounds

Fig. S3 plots the dynamics of group contributions over rounds of the catastrophe avoidance game across the four treatments. It can be seen from inspection of this figure that, with the exception of a trough in contributions at round 7, group contributions do not differ significantly over rounds in T_1 (Freidman, $\chi^2_{df=9} = 7.89$, $P = .545$), whereas group contributions decrease over rounds in T_2 (Freidman, $\chi^2_{df=9} = 23.89$, $P = .004$), with this decrease becoming more pronounced in the latter half of the game after the second set of proposals and pledges. Unlike T_2, group contributions in T_3 did not tail-off significantly over rounds (Freidman, $\chi^2_{df=9} = 5.90$, $P = .750$), indicating that the early-warning signal mid-game in this treatment helped to stabilise group contributions. The pattern of group contributions in T_4 is uniquely different from the remaining treatments. Although group contributions in this treatment decrease initially in the first half of the game, there is a punctuated peak in contributions on round 6 following the arrival of the early-warning signal, after which contributions decrease gradually, with a slight upturn in contributions on the final round (Freidman, $\chi^2_{df=9} = 15.61$, $P = .076$).

3.3 Proposals, pledges, and contributions

Fig. S4 shows the average group proposals and pledges on rounds 1 and 6, and group contributions (collapsed over the ten rounds) by treatment. Group proposals on rounds 1 and 6 hovered closely around the \$120 mark in all instances, confirming that this was the focal[6] threshold value in all treatments. In treatments T_1 and T_4, group pledges are generally consistent with the proposed group amounts, except that in T_4, on round 1 group pledges are lower than group proposals, whereas in treatments T_2

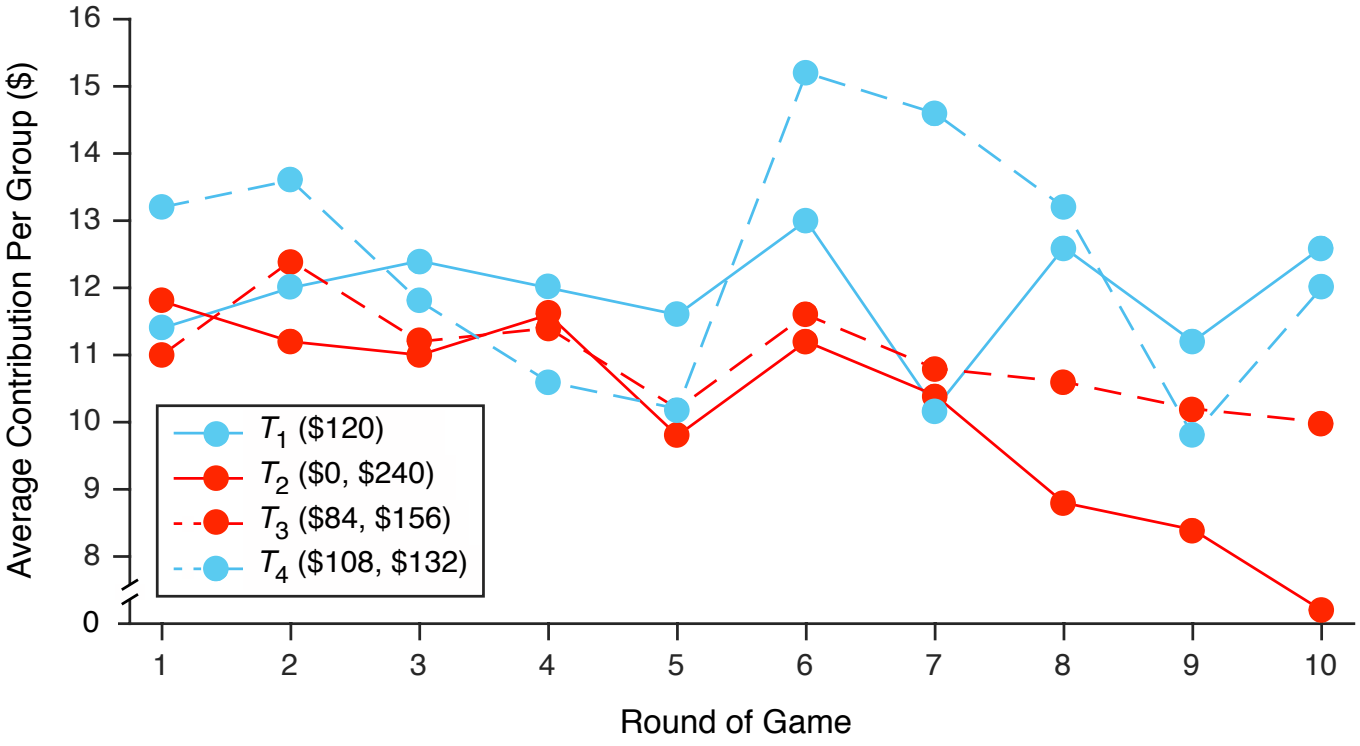

Figure S3 | Average group contributions over rounds of the catastrophe avoidance game as a function of treatment.

and T_3 group pledges are lower than group proposals on both rounds. There were no significant differences between treatments for group proposals on round 1 (Kruskal-Wallis, $\chi^2_{df=3} = 5.92$, $P = .115$) or round 6 (Kruskal-Wallis, $\chi^2_{df=3} = 1.98$, $P = .576$). However, collapsing across treatments, group proposals on round 6 (119.7 ± 1.63) were slightly, but significantly, lower than on round 1 (124.10 ± 2.12) (Wilcoxon Signed-Rank, W = 434.00, $P = .020$). There were no significant differences between treatments for group pledges on round 1 (Kruskal-Wallis, $\chi^2_{df=3} = 3.74$, $P = .291$) or round 6 (Kruskal-Wallis, $\chi^2_{df=3} = 4.31$, $P = .230$). Collapsing across treatments, there was no significant difference between group pledges on round 1 (115.70 ± 2.36) and round 6 (114.90 ± 2.12) (Wilcoxon Signed-Rank, W = 280.00, $P = .537$). Finally, there was a significant difference in group contributions as a function of treatment (Kruskal-Wallis, $\chi^2_{df=3} = 8.00$, $P = .046$). Contributions were significantly lower in T_2 than T_1 (Mann-Whitney, 80.50, $P = .023$), and although contributions did not differ significantly between T_2 and T_3 (Mann-Whitney, 39.00, $P = .427$), contributions were significantly higher in T_4 than T_2 (Mann-Whitney, 15.50, $P = .010$). Thus, a 90% reduction in threshold uncertainty succeeded in significantly increasing group contributions.

To determine if proposals and pledges were consequential with respect to actual contributions, we conducted a linear regression, with group contributions as the dependent variable and proposals and pledges on rounds 1 and 6 as predictors. The resulting model was significant, compared to a constant-only model, $F(4, 35) = 9.48$, $P < .001$. The results of the analysis are shown in Table S1, from which it can be seen that only group pledges on round 6 were a reliable signal of actual group contributions.

3.4 Economic preferences and contributions

To provide a further window into the factors that influenced individual contributions in the catastrophe avoidance game, participants completed an individual differences questionnaire at the end of the game which measured their risk, time, and social preferences.[20] Table S2 shows the average responses to each of the six economic preference items, which measured risk aversion, loss aversion, fairness, trust, altruism, and temporal discounting, as a function of the four treatments. Responses did not differ significantly across treatments for either of the economic preference items: (Kruskal-Wallis, $\chi^2_{df=3} = 4.55$, $P = .208$) for risk aversion, (Kruskal-Wallis, $\chi^2_{df=3} = 5.87$, $P = .118$) for loss aversion, (Kruskal-Wallis, $\chi^2_{df=3} = 0.56$, $P = .906$) for fairness, (Kruskal-Wallis, $\chi^2_{df=3} = 1.14$, $P = .768$) for trust, (Kruskal-Wallis, $\chi^2_{df=3} = 4.46$, $P = .216$) for altruism, and (Kruskal-Wallis, $\chi^2_{df=3} = 5.09$, $P = .165$) for temporal discounting. To examine whether economic preferences influenced player contributions in the catastrophe avoidance game, we conducted a linear regression with individual player contributions as the dependent measure and responses on each of the six economic preference items as the predictors. The model was not significant, relative to a constant-only model, $F(6, 233) = 1.66$, P

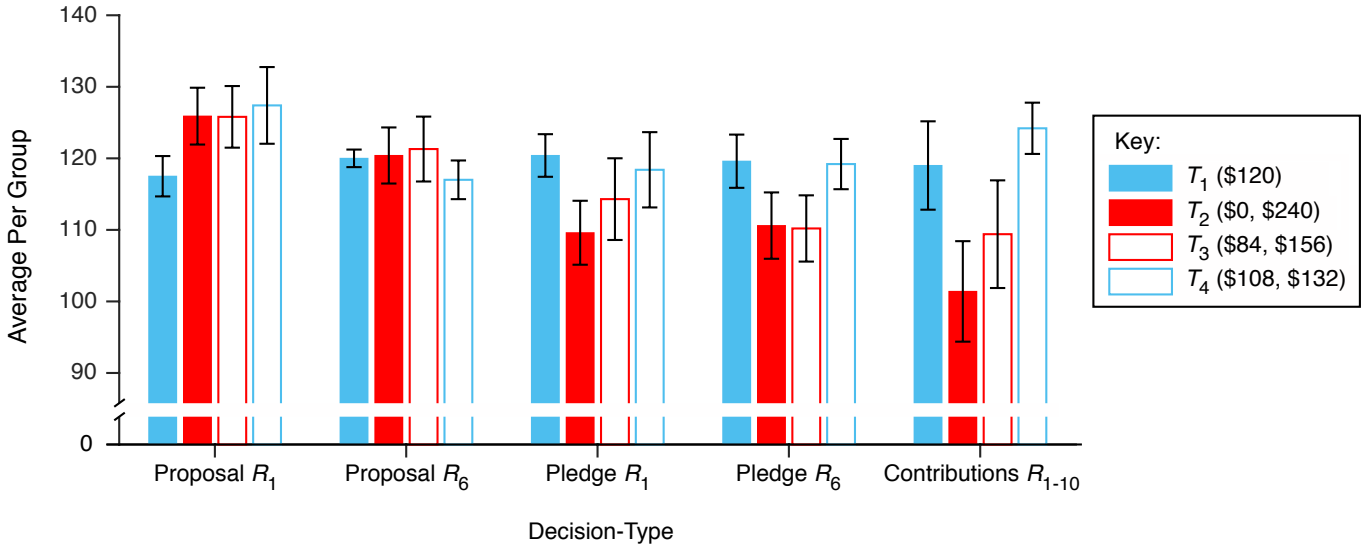

Figure S4 | Group proposals, pledges, and contributions as a function of treatment. Error bars represent standard errors.

Table S1 | Linear regression predicting group contributions

	Unstandardised *β*	Standard Error	Standardised *β*	*t*	*p*
Intercept	-42.47	38.21		-1.11	0.274
Proposal R_1	0.05	0.21	0.03	0.24	0.810
Proposal R_6	0.10	0.25	0.05	0.40	0.694
Pledge R_1	0.26	0.21	0.19	1.25	0.221
Pledge R_6	0.93	0.21	0.60	4.39	<.001

= .131. For completeness, Table S3 nevertheless summarises the results for each of the six predictors. Although risk aversion, loss aversion, fairness, and temporal discounting fell well short of statistical significance, trust (P = .061) and altruism (P = .137) were both close to being significant predictors. In brief, higher levels of self-reported dispositional trust and altruism were associated with higher contributions in the catastrophe avoidance game, although not statistically reliably so.

Table S2 | Mean responses on the post-game economic preferences questionnaire as a function of treatment

Construct	Question	T_1	T_2	T_3	T_4
Risk aversion	How do you see yourself: are you generally a person who is fully prepared to take risks or do you try to avoid taking risks?	6.12 (2.26)	5.48 (2.39)	5.88 (2.34)	5.30 (2.17)
Loss aversion	How well does the following statement describe you as a person? I generally hate to lose something more than I like to gain something.	6.68 (2.21)	6.10 (2.06)	5.88 (2.23)	6.40 (2.52)
Fairness	Please indicate your level of agreement with the following statement: when a group of people must work toward a common goal, it is important that each group member contributes an equal amount of effort.	8.22 (2.26)	8.12 (2.36)	8.47 (1.81)	8.32 (2.27)
Trust	How well does the following statement describe you as a person? As long as I am not convinced otherwise, I assume that people have only the best intentions.	6.05 (2.53)	5.57 (2.45)	5.63 (2.69)	5.75 (2.80)
Altruism	How willing are you to help others without expecting anything in return?	7.60 (1.98)	7.23 (1.92)	6.93 (2.02)	7.28 (2.12)
Temporal discounting	How willing are you to give up something today in order to benefit from doing so in the future?	7.85 (1.64)	7.48 (1.56)	7.22 (1.68)	7.28 (1.86)

All items required a response on an eleven point scale. For the risk aversion item, participants were asked to: "Please use a scale from 0 to 10, where 0 means you are completely unwilling to take risks and 10 means you are very willing to take risks"; for the loss aversion and trust items participants were asked to: "Please use a scale from 0 to 10, where 0 means does not describe me at all and 10 means describes me perfectly"; for the fairness item participants were asked to: "Please use a scale from 0 to 10, where 0 means strongly disagree and 10 means strongly agree"; for the altruism item participants were asked to: "Please use a scale from 0 to 10, where 0 means you are completely unwilling to help others and 10 means you are very willing to help others"; for the temporal discounting item participants were asked to: "Please use a scale from 0 to 10, where 0 means you are completely unwilling to give up something today and 10 means you are very willing to give up something today".

Table S3 | Linear regression predicting group contributions

	Unstandardised β	Standard Error	Standardised β	t	p
Intercept	14.23	2.97		4.81	<.001
Risk aversion	-0.04	0.18	-0.01	-0.21	0.834
Loss aversion	0.02	0.18	0.01	0.09	0.933
Fairness	0.07	0.20	0.03	0.37	0.711
Trust	0.32	0.17	0.13	1.88	0.061
Altruism	0.34	0.23	0.11	1.49	0.137
Temporal discounting	-0.02	0.24	-0.01	-0.08	0.940

References

1. Hurlstone, M. J., & Newell, B. R. Threshold uncertainty, early-warning signals, and the prevention of dangerous climate change. *Manuscript submitted for publication* (2019).
2. Barrett, S. & Dannenberg, A. Climate negotiations under scientific uncertainty. *Proc. Natl. Acad. Sci. USA* **109**(43), 17372–17376 (2012).
3. Barrett, S. & Dannenberg, A. Sensitivity of collective action to uncertainty about climate tipping points. *Nat. Clim. Change.* **4**, 36–39 (2014).
4. Dannenberg, A. *et al.* On the provision of public goods with probabilistic and ambiguous thresholds. *Environ. Resource. Econ.* **61**, 365–383 (2015).
5. Milinski, M., Sommerfeld R. D., Krambeck, H-J., Reed F. A., & Marotzke J. The collective-risk social dilemma and the prevention of simulated dangerous climate change. *P. Natl. Acad. Sci. USA* **105**, 2291–2294 (2008).
6. Schelling, T. C. *The strategy of conflict.* Harvard university press (1960).
7. Barrett, S. & Dannenberg, A. An experimental investigation into 'pledge and review' in climate negotiations. *Climatic Change* **138**, 339–351 (2016).
8. Tavoni, A., Dannenberg, A., Kallis, G., & Löschel, A. Inequality, communication, and the avoidance of disastrous climate change in a public goods game. *P. Natl. Acad. Sci. USA* **108**, 11825–11829 (2011).
9. Milinski, M. *et al.* Cooperative interaction of rich and poor can be catalyzed by intermediate climate targets. *Climatic Change* **109**, 807–814 (2011).
10. Burton-Chellew, M. N. *et al.* Combined inequality in wealth and risk leads to disaster in the climate change game. *Climatic Change* **120**, 815–830 (2013).
11. Jacquet, J. *et al.* Intra- and intergenerational discounting in the climate game. *Nat. Clim. Change* **3**, 1025–1028 (2013).
12. Barrett, S. Climate treaties and approaching catastrophes. *J. Environ. Econ. Manage.* **66**, 235–250 (2013).
13. Barrett, S., & Danneberg, A. Negotiating to avoid 'gradual' vs 'dangerous' climate change: An experimental test of two prisoners' dilemmas', in T. Cherry, J. Hovi and D. M. McEvoy (eds) *Towards a new climate agreement: Conflict, resolution, and governance*. London: Routledge, pp. 61–75 (2014).
14. Lenton, T. M. *et al.* Tipping Elements in the Earths Climate System. *Proc. Natl. Acad. Sci. USA* **105**(6), 1786–1793 (2008).
15. Lenton, T. Early warning of climate tipping points. *Nat. Clim. Change.* **1**, 201–209 (2011).
16. Lenton, T. M. *et al.* Early warning of climate tipping points from critical slowing down: Comparing methods to improve robustness. *Phil. Trans. R. Soc. A.* **370**, 1185–1204 (2012).
17. Scheffer, M. *et al.* Early-warning signals for critical transitions. *Nature* **461**, 53–59 (2009).
18. Scheffer, M. *et al.* Anticipating critical transitions. *Science* **338**, 344–338 (2012).
19. Lenton, T. M. Game theory: Tipping climate cooperation. *Nat. Clim. Chang.* **4**, 14–15 (2014).
20. Falk, A. *et al.* The preference survey module: A validated instrument for measuring risk, time, and social preferences. *SSRN Electronic Journal*, January (2016).