Sensitivity of collective action to early warning signals of shifting climate tipping points

Mark J. Hurlstone[1,*], Ben White[2], and Ben R. Newell[3,4]

[1]Lancaster University, Department of Psychology, Lancaster, LA1 4YW, UK
[2]University of Western Australia, School of Agriculture and Environment, Perth, 6039, WA, Australia
[3]UNSW, School of Psychology, Sydney, 2052, NSW, Australia
[4]UNSW, Institute for Climate Risk & Response, Sydney, 2052, NSW, Australia
[*]m.hurlstone@lancaster.ac.uk

The Paris Agreement aims to limit global warming to well below 2°C, and ideally to 1.5°C, above pre-industrial levels. The coexistence of these two policy benchmarks reflects deep uncertainties regarding the threshold for dangerous climate change. Theory and experimental evidence show that such uncertainty impedes collective action. Early warning signals offer the hope that, as proximity to a critical threshold increases, uncertainty about its location may shrink enough to trigger coordinated collective action to avoid it. Here we report an experiment in which groups coordinated on an uncertain threshold—akin to 2°C—before receiving an early warning signal of moderate or high precision signifying that the threshold had shifted closer than anticipated—akin to 1.5°C. Our results suggest that early warning signals may increase the probability of avoiding 2°C, but may not be sufficient to prevent overshooting 1.5°.

Keywords: Cooperation, Dangerous climate change, Early warning signals, Threshold uncertainty

The goal of the United Nations Framework Convention on Climate Change (UNFCCC), adopted in 1992, is to stabilise "greenhouse gas concentrations in the atmosphere at a level that would prevent dangerous anthropogenic interference with the climate system"[1]. The treaty therefore frames the climate negotiations around avoiding "dangerous" (abrupt and catastrophic) rather than "gradual" (smooth and continuous) climate change, which might occur if some critical atmospheric concentration level is exceeded. This framing is potentially advantageous—it may create a sense of fear and urgency that compels countries to reduce their emissions to avoid crossing the critical threshold—but to be effective, the threshold needs to be scientifically credible[2]. However, the UNFCCC did not define a threshold for dangerous climate change. In 2009, the signatories of the Copenhagen Accord agreed on a benchmark, namely that in accordance with "the scientific view the increase in global temperature should be below 2 degrees Celsisus"[3]. Yet, only a year later in Cancún, the negotiators considered the scientific need to strengthen this goal to 1.5°C[4]. This uncertainty is reflected in the Paris Agreement, which both reaffirms the 2°C target and underscores the desirability of "pursuing efforts to limit the temperature increase to 1.5°C"[5]. The coexistence of these two policy benchmarks reflects deep and persistent uncertainties in the climate science over the location of the critical threshold for dangerous climate change[6–8].

The consequences of threshold uncertainty for the climate negotiations have been studied in the laboratory using threshold experiments[9–11]. In these games, groups of players must cooperate by investing money from a personal endowment into hypothetical emission abatement to avoid crossing a dangerous threshold, which, if breached, triggers catastrophic economic losses for all. Theory predicts that when the threshold is certain or uncertainty is very small, groups should be able to coordinate their efforts to stay on the safe side of the dangerous threshold, whereas when uncertainty is large, cooperation should break down, with groups straying into the danger zone[12]. Experimental evidence confirms these predictions: compared to a certainty situation, threshold uncertainty erodes group contributions and increases the probability of crossing the threshold[13–15], unless the uncertainty is confined to a narrow range[14]. The negative effect of threshold uncertainty on cooperation has been observed in both one-shot[13–15] and iterated[16–18] games; in the presence[13,14,17,18] and absence[16] of communication between players; with symmetric and asymmetric player endowments[16]; and when the threshold is ambiguous rather than merely uncertain[17]. The presence of an uncertain threshold nevertheless facilitates cooperation compared to a no-threshold scenario[15]. This suggests that the framing of the climate negotiations in terms of avoiding "dangerous" instead of "gradual" climate change has been beneficial[15]—an uncertain dangerous threshold may motivate countries to reduce their emissions more than if no threshold existed at all. However, it may be insufficient to prevent countries from averting catastrophe.

An obvious implication of these findings is that if uncertainty around the threshold for dangerous climate change could be reduced, this might trigger collective action to avoid it. A growing body of work has shown that uncertainty about the location of some dangerous tipping points can be reduced through the detection of early warning signals of approaching climate

Table 1. Overview of the design of the experiment.

Treatment	Uncertainty	Early warning	Rounds	$\bar{Q}$	ε	$Q_{\min}$	$Q_{\max}$	N participants
Low-threshold baseline								
100	No	No	1–10	100	0	100	100	25 × 5 = 125
50/150	Yes	No	1–10	100	50	50	150	25 × 5 = 125
High-threshold baseline								
150	No	No	1–10	150	0	150	150	25 × 5 = 125
100/200	Yes	No	1–10	150	50	100	200	25 × 5 = 125
Early warning signal								
135/165	Yes	Yes	1–5	100	50	50	150	25 × 5 = 125
			6–10	150	15	135	165	
145/155	Yes	Yes	1–5	100	50	50	150	25 × 5 = 125
			6–10	150	5	145	155	

Note: $\bar{Q}$ = mean of the threshold distribution; ε = half-width of the threshold distribution; $Q_{\min}$ = lower threshold bound; $Q_{\max}$ = upper threshold bound.

transitions[19–23]. For example, strong positive feedback in the internal dynamics of the climate system, or generic statistical indicators of loss of system resilience, may signal that a tipping point is near[21].

In a recent iterated threshold experiment, we tested this idea directly[18]. The experiment included a certainty treatment, where the threshold was fixed, and an uncertainty treatment, with wide uncertainty about the threshold. Additionally, there were two early-warning treatments. These were identical to the uncertainty treatment in the first half of the game. However, in the second half, both treatments received a surprise mid-game warning indicating that the uncertainty around the expected value of the threshold had been reduced. In the wide-warning treatment, uncertainty was reduced by 70%, while in the narrow-warning treatment it was reduced by 90%. Consistent with prior studies, uncertainty about the threshold reduced group contributions and the probability of avoiding catastrophe, compared to the certainty case. The wide-warning had no effect. By contrast, the narrow warning triggered a regime shift: contributions in the second half of the game increased sharply, with total contributions rising to a level comparable to the certainty treatment, and the probability of avoiding catastrophe more than doubled compared to the uncertainty case.

Generalising to the real game of climate change, these findings suggest that if countries are coordinating on an uncertain 2°C target, an early warning confirming this threshold may increase the probability of avoiding it, provided uncertainty is reduced to within a very narrow range. Yet early warnings may do more than reduce uncertainty—they may also reveal that the critical threshold is closer than anticipated. For example, a warning might indicate that the threshold lies nearer to 1.5°C. Here, we report a threshold experiment designed to test whether early warning signals that not only reduce uncertainty but also shift the expected threshold value can trigger coordinated action to avoid crossing it.

Current study

Our experiment involved 750 participants, randomly assigned to five-player groups, who played an iterated threshold game with communication[17,18,24]. At the start of the game, each player was given a 40-token endowment. On each of ten rounds, players decided whether to contribute 0, 2, or 4 tokens into a collective damage-avoidance account. The players knew that if the total amount contributed by the end of the game equalled or exceeded a threshold amount, then they would get to keep their remaining endowment; otherwise they would lose 90% of it. Before the contribution decisions in rounds 1 and 6, each player submitted two non-binding communications: (*i*) a proposal of how many tokens the group should collectively contribute over all 10 rounds, and (*ii*) a pledge of how many tokens they personally intended to contribute toward the agreed goal.

The experiment comprised six treatments, each with 25 groups (Table 1). Two were low-threshold baseline treatments, with a mean threshold of 100 tokens. In treatment 100, the threshold was certain: groups were told that it was fixed at 100 tokens. In treatment 50/150, the threshold was uncertain: groups were told it would be a random value between 50 and 150 tokens. Two further treatments, 150 and 100/200, were high-threshold baselines, mirroring the certainty and uncertainty structure of the low-threshold treatments but with a mean threshold of 150 tokens. The final two treatments, 135/165 and 145/155, were the early-warning treatments. In the first half of the game, these treatments were functionally identical to treatment 50/150. However, at the mid-point of the game, before the second set of non-binding communications, groups received a surprise warning indicating that the threshold information had changed. In treatment 135/165, groups were told that the threshold would now be a random amount between 135 and 165 tokens (a 70% reduction in uncertainty, narrowing it to within ±15 tokens of

Table 2. Cooperative and Nash equilibria for total contributions (rounds 1–10) and expected contributions in the first and second halves of the game (rounds 1–5 and 6–10) for each treatment, with predicted success at the 100- and 150-token thresholds.

		Group contributions						Threshold reached?			
		Total		Rounds 1–5		Rounds 6–10		Cooperative		Nash	
Treatment	Rounds	Cooperative	Nash	Cooperative	Nash	Cooperative	Nash	100	150	100	150
Low-threshold baseline											
100	1–10	100.00	100.00	50.00	50.00	50.00	50.00	Yes	No	Yes	No
50/150	1–10	119.44	65.74	59.72	32.87	59.72	32.87	Yes	No	No	No
High-threshold baseline											
150	1–10	150.00	150.00	75.00	75.00	75.00	75.00	Yes	Yes	Yes	Yes
100/200	1–10	144.44	107.41	72.22	53.70	72.22	53.70	Yes	No	Yes	No
Early warning signal											
135/165	1–5	119.44	65.74	59.72	32.87	—	—	—	—	—	—
	6–10	165.00	143.06	—	—	100.00	100.00	Yes	Yes	Yes	No
145/155	1–5	119.44	65.74	59.72	32.87	—	—	—	—	—	—
	6–10	155.00	153.24	—	—	95.28	100.00	Yes	Yes	Yes	Yes

Note: Predictions for rounds 6–10 are capped at 100 tokens to reflect the maximum feasible group contribution in this period. The corresponding uncapped second-half predictions are 105.28 (cooperative) and 110.19 (Nash) for treatment 135/165; and 120.37 (Nash) for treatment 145/155.

the threshold location, but simultaneously increasing its expected value from 100 to 150 tokens). In treatment 145/155, groups were told that the threshold would now be a random amount between 145 and 155 tokens (a 90% reduction in uncertainty, narrowing it to within ±5 tokens of the threshold location, but simultaneously increasing its expected value from 100 to 150 tokens). In our experiment, we tentatively take the 100- and 150-token mean threshold values to represent the 2°C and 1.5°C policy benchmarks in the Paris Agreement.

The low-threshold baselines implement the certainty and uncertainty treatments typical of threshold uncertainty experiments, where players must contribute on average half of their endowment to reach the mean of the threshold distribution[16–18]. The certainty versus uncertainty comparison is included to replicate the established finding that uncertainty about the threshold undermines cooperation relative to the certainty situation. The high-threshold baselines are novel. They examine threshold uncertainty in a more demanding setting, where players must contribute on average three-quarters of their endowment to reach the mean of the threshold distribution. Two sets of baselines are required to evaluate the early-warning treatments (135/165 and 145/155) because the mean of the threshold distribution changes within these treatments—from 100 tokens in the first half of the game to 150 tokens in the second half.

By comparing treatments 135/165 and 145/155 with treatment 50/150, we ask whether groups that initially coordinate on an uncertain 100-token threshold, but are later told the threshold is closer to 150 tokens, are more or less likely to reach the 100-token threshold than groups that never receive a warning. This is analogous to asking whether countries that begin by coordinating on an uncertain 2°C target, but are subsequently informed that the true threshold is closer to 1.5°C, are more or less likely to avoid crossing 2°C than countries that received no such warning.

By comparing treatments 135/165 and 145/155 with treatment 100/200, we ask whether groups that initially coordinate on an uncertain 100-token threshold, but are later told the threshold is closer to 150 tokens, are more or less likely to reach the 150-token threshold than groups that coordinated on this higher uncertain threshold from the outset. This is analogous to asking whether countries that begin by coordinating on an uncertain 2°C target, but are subsequently informed that the true threshold is closer to 1.5°C, are more or less likely to avoid crossing 1.5°C than countries that were coordinating on an uncertain 1.5°C target from the outset.

The two early-warning treatments allow us to determine whether the answer to the above questions depend on the precision with which early warnings pinpoint the new threshold. Treatment 135/165 provided a moderately precise signal, whereas treatment 145/155 provided a highly precise signal. Including both allowed us to test whether the effectiveness of an early warning depends on its precision, as per our earlier study[18].

Game-theoretic predictions

Our predictions are based on a game-theoretic model that specifies the cooperative and Nash equilibria for each treatment in terms of total contributions, contributions in the first and second halves of the game, and whether the 100- and 150-token thresholds to avoid catastrophe are successfully reached (Table 2).

For the low-threshold baseline treatments, under certainty (100), the cooperative and Nash equilibria coincide at 100 tokens, ensuring the threshold is reached. Under uncertainty (50/150), the cooperative equilibrium (119.44) exceeds the Nash equilibrium (65.74). The cooperative prediction implies success at the 100-token threshold, whereas the Nash prediction falls short, meaning success depends on whether groups coordinate on the cooperative or Nash equilibrium. Neither equilibrium reaches 150 tokens, whether the threshold is certain or uncertain.

For the high-threshold baseline treatments, under certainty (150) the cooperative and Nash equilibria coincide at 150 tokens, guaranteeing that both the 100- and 150-token thresholds are reached. Under uncertainty (100/200), the cooperative equilibrium (144.44) exceeds the Nash equilibrium (107.41). Both equilibria surpass 100 tokens but fall short of 150.

Since the threshold does not change in these treatments, contributions are expected to be evenly distributed over the first and second halves of the game. By contrast, in the early-warning treatments, the mid-game warnings trigger a regime shift in the second half. In rounds 1–5, the predictions mirror treatment 50/150. After the warning, predicted second-half contributions rise sharply: in treatment 135/165, both equilibria predict 100 tokens, while in treatment 145/155 the cooperative equilibrium predicts 95.28 tokens and the Nash equilibrium 100.00. Accordingly, total contributions rise toward the new mean threshold of 150 tokens in both treatments. In treatment 135/165, the cooperative equilibrium (165.00) exceeds the Nash equilibrium (143.06), implying success at the 150-token threshold only if groups coordinate on the cooperative equilibrium. In treatment 145/155, the cooperative (155.00) and Nash (153.24) equilibria almost converge, and both predict surpassing 150 tokens.

In brief, certainty should guarantee avoiding catastrophe at the relevant threshold: treatment 100 should succeed at 100 tokens, while treatment 150 should succeed at both 100 and 150. Under uncertainty, treatment 50/150 can avoid catastrophe at the 100-token threshold only if groups coordinate on the cooperative equilibrium, whereas treatment 100/200 should avoid catastrophe at 100 tokens but not at 150. Early warning signals should elevate contributions toward 150: in treatment 135/165, avoiding catastrophe at this higher benchmark is possible only under the cooperative equilibrium, while in treatment 145/155 both equilibria imply that catastrophe should be avoided.

Results

We use robust statistics and nonparametric tests to compensate for non-normal responses and outliers typical of data from threshold experiments. For all analyses, the basic statistical unit is the group. Although our primary focus is on the effect of treatment on group contributions and the percentage of groups that would have succeeded in avoiding catastrophe at the 100- and 150-token thresholds, we also analyse the effect of treatment on group proposal and pledge amounts over time, and the relationship of proposals and pledges to contributions.

Contributions and equilibria

We begin by examining group contributions—both in total and separately for the first and second halves of the game—and their relation to the cooperative and Nash equilibria (Supplementary Analyses: Equilibrium Comparisons). To assess the impact of the mid-game warnings, we conducted two sets of analyses: (*i*) comparing the early-warning treatments (135/165, 145/155) with the low-threshold baseline treatments (100, 50/150; *low-threshold comparison*); and (*ii*) comparing the early-warning treatments with the high-threshold baseline treatments (150, 100/200; *high-threshold comparison*).

Starting with total contributions (Fig. 1a), contributions in treatments 100 and 150 aligned with the cooperative and Nash equilibria, which coincide, whereas contributions in treatments 50/150 and 100/200 were uniquely consistent with the cooperative equilibria. Contributions in treatments 135/165 and 145/155 were consistent with the Nash equilibria associated with the new mid-game threshold ranges.

For the low-threshold comparison, contributions differed significantly by treatment (Kruskal-Wallis, $\chi^2_{df=3} = 52.03$, $p < .001$). There was no uncertainty effect: contributions did not differ significantly between treatments 100 and 50/150 (Mann-Whitney $U = 258.00$, $p = .292$). However, compared to treatment 50/150, contributions were significantly higher in treatment 135/165 (Mann-Whitney $U = 104.50$, $p < .001$) and treatment 145/155 (Mann-Whitney $U = 43.00$, $p < .001$). For the high-threshold comparison, contributions again differed significantly by treatment (Kruskal-Wallis, $\chi^2_{df=3} = 8.55$, $p = .036$). Here, an uncertainty effect was observed: contributions were significantly lower in treatment 100/200 than in treatment 150 (Mann-Whitney $U = 188.00$, $p = .016$). By contrast, contributions in treatments 135/165 and 145/155 did not differ significantly from treatment 100/200 (Mann-Whitney $U = 305.00$, $p = .900$, and $U = 240.00$, $p = .162$, respectively), despite the visible trend toward higher contributions in treatment 145/155.

Turning to the first half of the game (Fig. 1b), contributions in treatments 100 and 150 again aligned with the cooperative and Nash equilibria, while contributions in treatments 50/150 and 100/200 were uniquely consistent with the cooperative equilibria. Contributions in treatments 135/165 and 145/155 were consistent with the cooperative equilibria associated with the original threshold ranges.

For the low-threshold comparison, contributions did not differ significantly by treatment (Kruskal-Wallis, $\chi^2_{df=3} = 0.59$, $p = .898$). This confirms that prior to the mid-game warnings, contributions in treatments 135/165 and 145/155 did not differ from

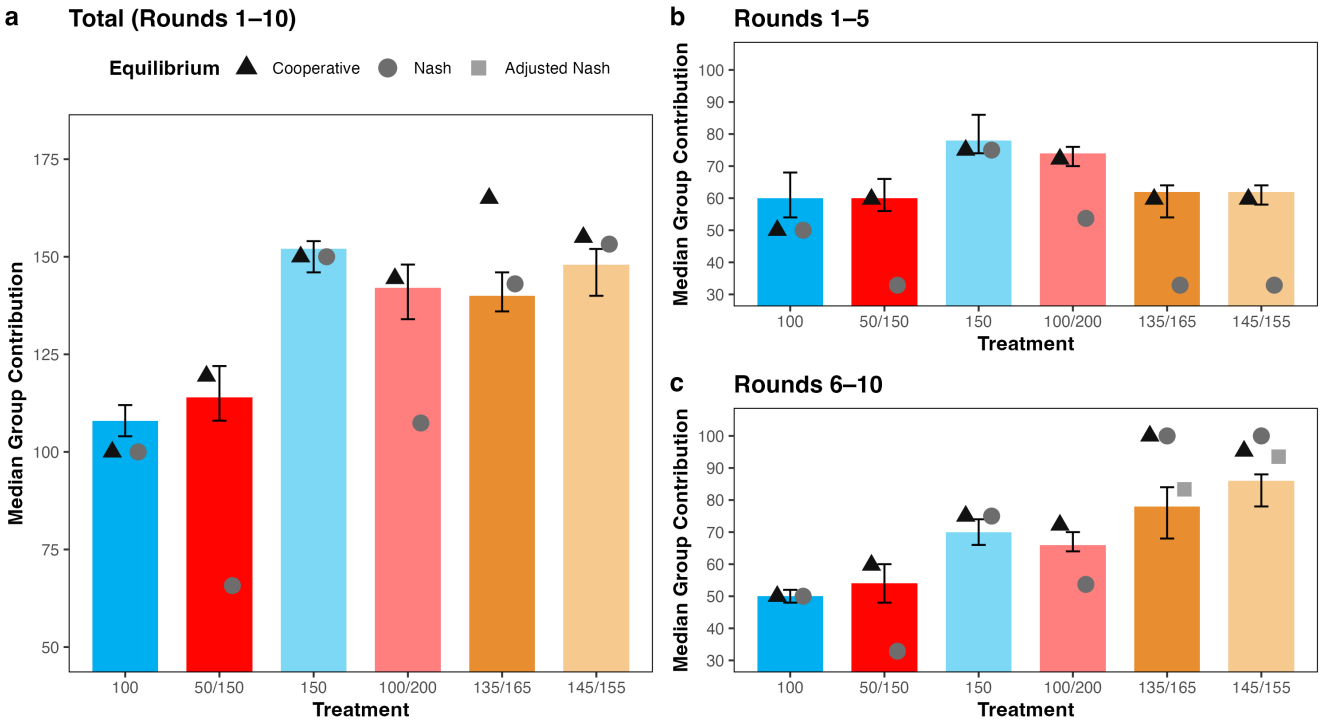

Figure 1. Median group contributions versus predicted values by treatment. a, Total (rounds 1–10), **b,** rounds 1–5, and **c,** rounds 6–10. Bars show the median group contribution within each treatment; error bars are 95% confidence intervals estimated using a percentile bootstrap (2,000 resamples; resampling with replacement). Overlaid markers indicate model predictions: cooperative equilibrium (triangle), Nash equilibrium (circle), and adjusted Nash equilibrium (square; shown in **c** for treatments 135/165 and 145/155). Predictions for treatments 135/165 and 145/155 in **a** reflect the narrowed threshold range announced mid-game. The adjusted Nash equilibrium in **c** assumes groups coordinated on the cooperative equilibrium in rounds 1–5 associated with the original threshold range. *Y*-axis limits differ by panel for visual clarity.

those in treatment 50/150, as expected because the treatments were functionally identical in this phase of the game. For the high-threshold comparison, contributions differed significantly by treatment (Kruskal-Wallis, $\chi^2_{df=3} = 30.95$, $p < .001$). An uncertainty effect emerged: contributions were significantly lower in treatment 100/200 than in treatment 150 (Mann-Whitney $U = 209.50$, $p = .046$). Moreover, compared to treatment 100/200, contributions were significantly lower in treatment 135/165 (Mann-Whitney $U = 128.5$, $p < .001$) and treatment 145/155 (Mann-Whitney $U = 150.00$, $p = .002$). These differences were expected, given that groups in treatment 100/200 were coordinating on a higher threshold and the mid-game warnings in treatments 135/165 and 145/155 had not yet been announced.

Finally, we consider the second half of the game (Fig. 1c), after the announcement of the mid-game warnings in treatments 135/165 and 145/155. Once again, contributions in treatments 100 and 150 aligned with the cooperative and Nash equilibria, while contributions in treatments 50/150 and 100/200 were uniquely consistent with the cooperative equilibria. Contributions in treatments 135/165 and 145/155 most closely matched the Nash equilibria associated with the newly announced threshold ranges, adjusted based on the assumption that in the first half of the game, groups coordinated on the cooperative equilibrium associated with the original threshold ranges (see 'Adjusted Nash' in Fig. 1c).

For the low-threshold comparison, contributions differed significantly by treatment (Kruskal-Wallis, $\chi^2_{df=3} = 61.01$, $p < .001$). There was no uncertainty effect: contributions were slightly higher in treatment 50/150 than treatment 100, although the difference fell just short of conventional significance (Mann-Whitney $U = 221.50$, $p < .078$). By contrast, compared to treatment 50/150, contributions were significantly higher in treatment 135/165 (Mann-Whitney $U = 51.50$, $p < .001$) and treatment 145/155 (Mann-Whitney $U = 31.50$, $p < .001$). For the high-threshold comparison, contributions again differed significantly by treatment (Kruskal-Wallis, $\chi^2_{df=3} = 28.90$, $p < .001$). There was no uncertainty effect: contributions did not differ significantly between treatments 150 and 100/200 (Mann-Whitney $U = 251.50$, $p = .239$). Critically, however, compared to treatment 100/200, contributions were significantly higher in treatment 135/165 (Mann-Whitney $U = 138.50$, $p < .001$) and treatment 145/155 (Mann-Whitney $U = 84.50$, $p < .001$). These results confirm that the mid-game announcements in the early warning treatments significantly increased group contributions relative to both the low- and high-threshold uncertainty

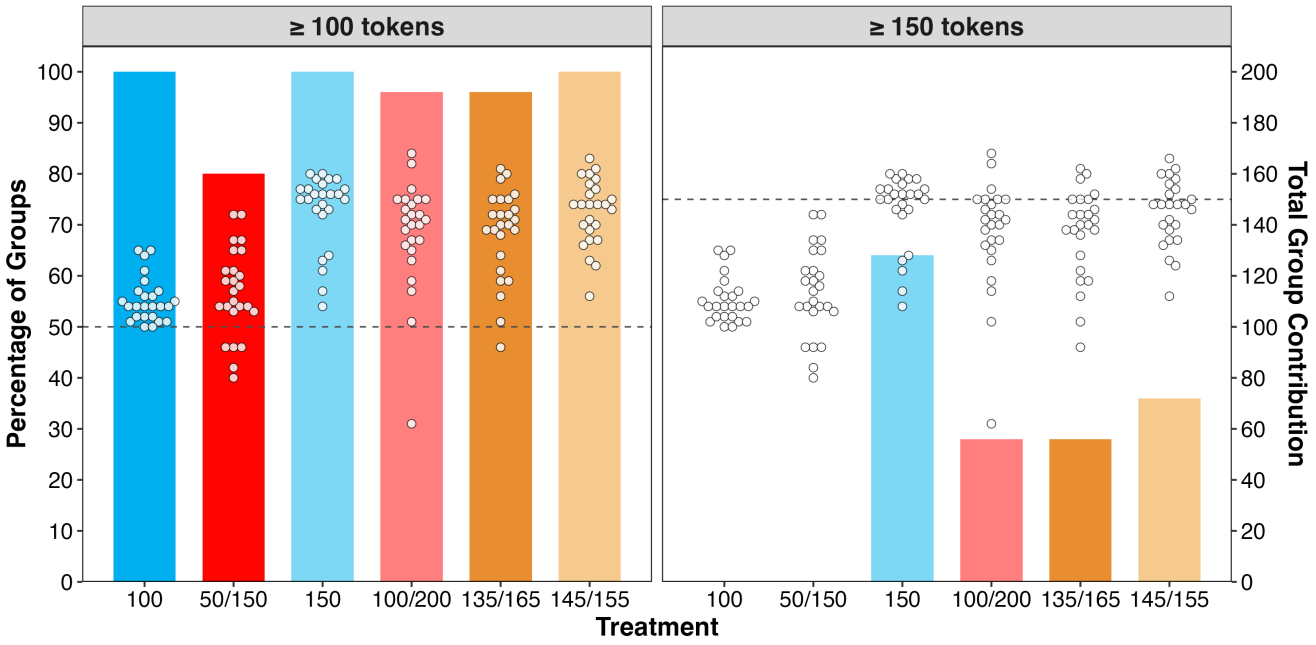

Figure 2. Success rates at critical threshold values by treatment. Bars show the percentage of groups reaching 100 tokens (left) and 150 tokens (right) by treatment (primary *y*-axis). Points show total contributions per group across rounds 1–10 (secondary *y*-axis). Hatched reference lines indicate 100 token (left) and 150 token (right) contribution levels.

treatments in the second half of the game.

Percentage of groups reaching 100-/150-tokens

We now examine the percentage of groups that would have succeeded in averting catastrophe at the 100- and 150-token thresholds (Fig. 2). These thresholds provide natural benchmarks for evaluating group success. Specifically, 100 tokens is the actual threshold in treatment 100 and the expected threshold in treatment 50/150 (and in treatments 135/165 and 145/155 for the first half of the game). Similarly, 150 tokens is the actual threshold in treatment 150 and the expected threshold in treatment 100/200 (and in treatments 135/165 and 145/155 for the second half of the game).

Starting with the 100-token threshold, for the low-threshold comparison, the percentage of groups reaching the threshold differed significantly across treatments (Fisher–Freeman–Halton, $p = .014$). There was a trend toward an uncertainty effect: all groups in treatment 100 succeeded, whereas only 80% of groups in treatment 50/150 did so, although the comparison fell just short of conventional significance (Fisher exact, $p = .050$). In treatments 135/165 and 145/155, the percentage of groups reaching the threshold increased to 90% and 100%, respectively. The comparison between treatments 50/150 and 135/165 was not significant (Fisher exact, $p = .190$), whereas the comparison between treatments 50/150 and 145/155 was marginal (Fisher exact, $p = .050$). For the high-threshold comparison, no significant differences between treatments emerged (Fisher–Freeman–Halton, $p = 1.000$): nearly all groups in treatments 150, 100/200, 135/165, and 145/155 contributed at least 100 tokens.

Turning to the 150-token threshold, we consider only the high-threshold comparison. There was a significant effect of treatment on the percentage of groups reaching the threshold (Fisher–Freeman–Halton, $p = .031$). There was a reliable uncertainty effect: 64% of groups reached the threshold in treatment 150, compared with only 28% in treatment 100/200—a significant difference (Fisher exact, $p = .022$). However, relative to treatment 50/150, the mid-game warnings did not significantly increase success rates: 28% in treatment 135/165 (Fisher exact, $p = 1.000$) and 36% in treatment 145/155 (Fisher exact, $p = .762$).

In treatment 145/155, there were a cluster of groups that fell marginally short of reaching the 150-token threshold. Based on this observation, we undertook a near-miss analysis of success rates at 148 tokens. The effect of treatment increased in significance compared to the analysis at 150 tokens (Fisher–Freeman–Halton, $p = .012$). The percentage of groups that reached the threshold increased in three of the four treatments: from 64% to 68% in treatment 150; from 28% to 32% in treatment 100/200; and, importantly, from 36% to 56% in treatment 145/155. The success rate in treatment 135/165 remained at 28%. Notably, the increase from 32% in treatment 100/200 to 56% in treatment 145/155 corresponds to a 24-percentage-point gain (a

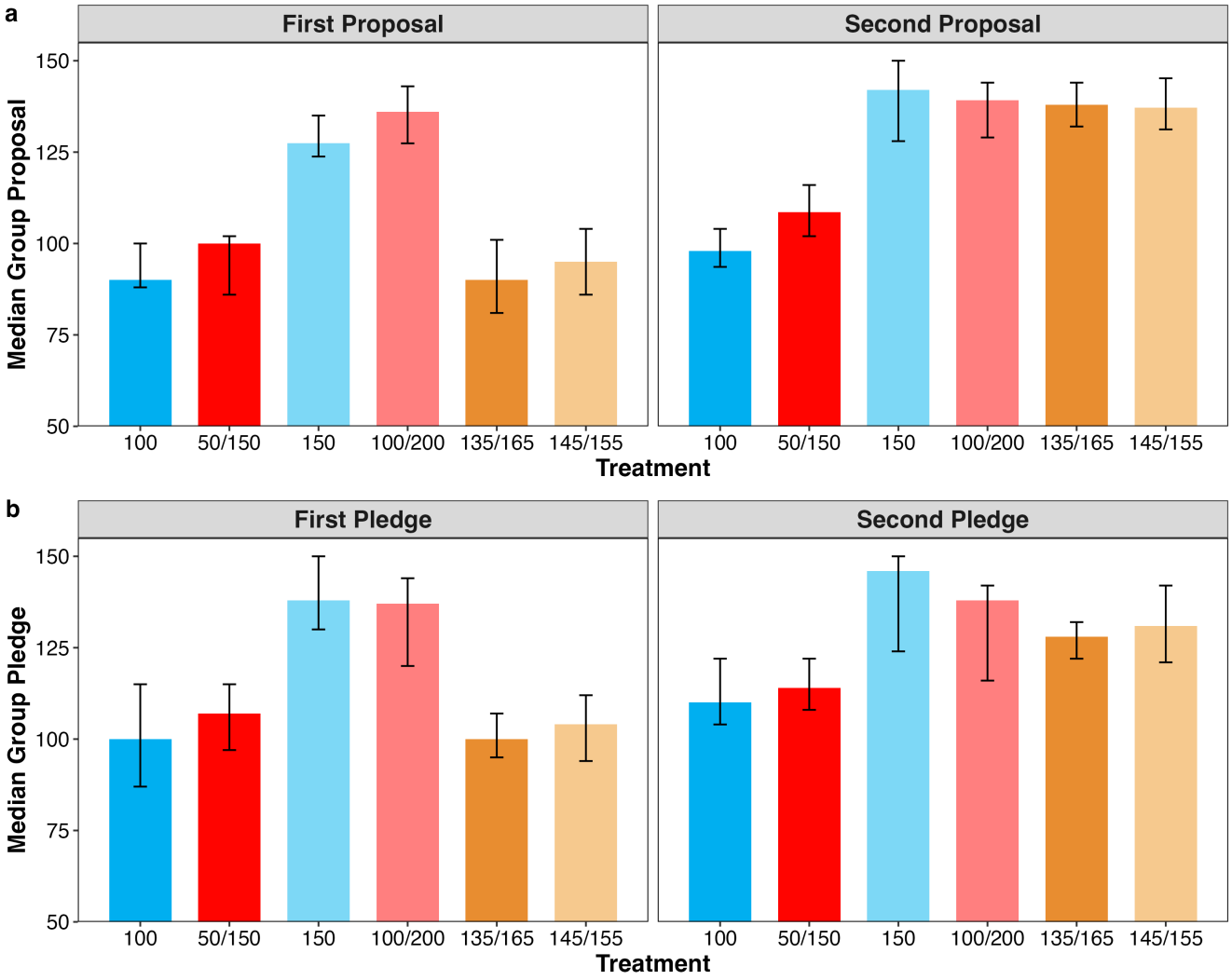

Figure 3. Group proposals and pledges by treatment. a, Median average first and second group proposal amounts. **b,** Median first and second total group pledge amounts. Error bars represent 95% confidence intervals estimated using a percentile bootstrap (2,000 resamples; resampling with replacement).

75% relative increase). The comparison between treatment 150 and 100/200 was significant (Fisher exact, $p = .023$), whereas the comparison between treatment 100/200 and treatment 135/165 remained nonsignificant (Fisher exact, $p = 1.000$). However, the comparison between treatment 100/200 and treatment 145/155 now approached, but did not reach, conventional significance (Fisher exact, $p = .154$).

Proposals, pledges, and their relation to contributions

The impact of the mid-game warnings on group contributions was mirrored broadly in average group proposals (Fig. 3a) and total group pledges (Fig. 3b). In the first half of the game, first proposals and pledges were close to 100 tokens in treatments 100, 50/150, 135/165, and 145/155, while they were higher in treatments 150 and 100/200 but still fell short of 150. Following the mid-game warnings, second proposals and pledges rose markedly in treatments 135/165 and 145/155, reaching levels comparable to those in treatments 150 and 100/200. These trends were statistically confirmed via analyses of group differences (Supplementary Analyses: Proposals and Pledges), using the same low-threshold and high-threshold comparisons reported in the preceding sections.

The degree to which proposals and pledges served as reliable signals of total contributions varied according to the timing of these non-binding communications (Fig. 4). First proposal and pledge amounts were generally weak and inconsistent predictors of total contributions across treatments. By contrast, second proposal and pledge amounts were consistently and positively associated with total contributions across treatments. The one exception was the absence of an association between second

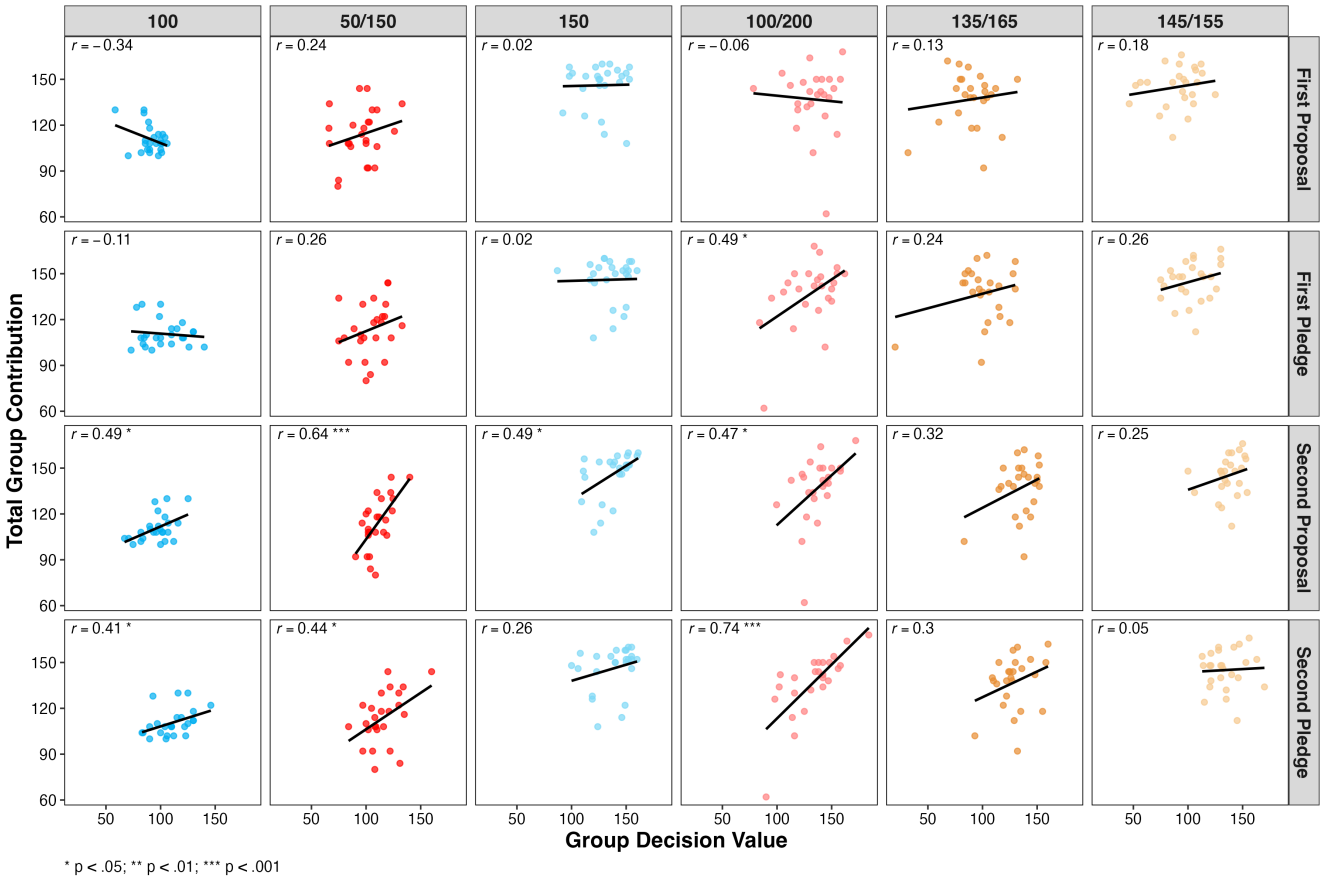

Figure 4. Group proposals, pledges, and contributions by treatment. Each panel relates total group contributions (*y*-axis) to the corresponding group decision value (*x*-axis): average first proposals (first row), total first pledges (second row), average second proposals (third row), and total second pledges (fourth row). Columns denote treatments. Each point corresponds to a group and the black lines are ordinary least-squares fits. Panel annotations report Pearson's *r* within each treatment, with significance indicated by stars (see key at bottom of figure).

pledges and contributions in treatment 145/155, which reflected the restricted range of pledges and contributions resulting from the extreme narrowing of the threshold range mid-game in this treatment. A linear regression analysis confirmed these trends (Supplementary Analyses: Relationship of Proposals and Pledges to Contributions): second proposals and pledges were significant positive predictors of contributions, whereas first proposals and pledges were not.

Discussion

The goal of the Paris Agreement is to limit global warming to well below 2°C, and ideally to 1.5°C, above pre-industrial levels[5]. The coexistence of these two policy benchmarks implies that the threshold for dangerous climate change is uncertain, and the scientific literature confirms this[6–8]. Theory and experimental evidence converge in showing that collective action is feasible when the threshold is certain or nearly so, but not when uncertainty is large[12–15]. The science of early warning signals[19–23] offers the hope that as proximity to a dangerous tipping point increases, uncertainty about its location might be reduced sufficiently to trigger pre-emptive collective action to avoid it. Recent experimental work supports this possibility: early warnings that reduced large uncertainty around an expected threshold increased group contributions and the likelihood of avoiding catastrophe, but only when the warning was highly precise[18]. Here, we examined a more complex scenario in which early warnings not only reduce uncertainty but also signal that the threshold lies closer than anticipated—analogous to countries coordinating on an uncertain 2°C target and then receiving evidence, with moderate or high precision, that the threshold is nearer 1.5°C.

Our theory predicted that under certainty, groups would succeed in avoiding catastrophe at the relevant threshold, whereas with an uncertain threshold, cooperation would be harder to mobilise. The results broadly supported these predictions, with two

important exceptions. First, although groups always succeeded under certainty when the threshold was low, only 64% did so when it was high. To our knowledge, this is the first experimental evidence that certainty alone may be insufficient to secure cooperation when the threshold requires a higher collective contribution. Second, the impact of uncertainty varied by threshold level: at the high threshold, uncertainty reduced both contributions and success rates, while at the low threshold, contributions were unaffected but success was marginally less likely. The latter result is contrary to previous studies showing that uncertainty erodes contributions[13–18]. The most likely explanation for this discrepancy is that our experiment used a narrower uncertainty range compared to previous iterated threshold experiments with communication[17,18]—in those studies, our theoretical model predicts an uncertainty effect for both the cooperative and Nash equilibria and contributions under uncertainty are consistent with the cooperative equilibrium prediction[18]. Here, groups also coordinated on the cooperative equilibrium, but our theoretical model did not predict an uncertainty effect at this benchmark, it only did so for the Nash equilibrium. Taken together, however, these outcomes align with the predicted equilibria. In the four baseline treatments, groups coordinated consistently on the cooperative equilibrium under uncertainty, and on both the cooperative and Nash equilibria under certainty.

Our theory further predicted that early warnings would catalyse contributions toward the revised, higher threshold. When the warning was moderately precise, success was possible only under the cooperative equilibrium, whereas when it was highly precise, success was expected at both equilibria. Thus, the more precise warning was expected to elicit the stronger behavioural response. The results partially confirmed these predictions. Early warnings triggered a regime shift: groups initially coordinated on the cooperative equilibrium associated with the original threshold, but after the warning shifted toward the Nash equilibrium associated with the revised, higher threshold. Consequently, early warnings raised contributions to levels comparable with the high-threshold baselines, but they did not increase the probability of avoiding catastrophe relative to groups facing the uncertain high threshold from the outset. One important caveat is that with a highly precise warning, groups that failed did so only marginally—when success is defined within one contribution of the high threshold, success rates matched those observed under certainty.

If we tentatively take the low and high thresholds in our experiment to represent the 2°C and 1.5°C policy benchmarks in the Paris Agreement, then our results suggest several implications for the climate negotiations. First, a highly precise early warning signal could increase the probability of countries avoiding crossing the 2°C target, but contributions may still fall short of what is required to stay below 1.5°C. This implies that early warning systems alone will be insufficient to bridge the gap in collective action, and that complementary mechanisms—such as reward[25], punishment[25,26], or reputation[27,28]—will be needed to mobilise contributions at the higher level. This is especially true given our finding that at the higher threshold level, the stabilising effect of certainty on cooperation greatly diminishes. Thus, even if an early warning signal can raise cooperation to a level comparable to that when the threshold is certain, it would still leave a high probability of disaster occurring. Second, the effectiveness of early warnings depends critically on their precision. In our study, uncertainty had to be reduced to within about 10% of the true threshold location to elicit a robust cooperative response, consistent with our earlier study[18]. Whether such precision can be achieved in practice is unclear[29]. It may only be possible once the climate system is already very close to a critical threshold, by which time there may be too little time left to take evasive action to avoid overstepping it[30]. Finally, if early warning signals of sufficient precision can be detected, our findings suggest they should be coupled with a requirement for countries to submit new Nationally Determined Contributions[5]. In this, and our earlier experiment[18], the most recent proposals and pledges provided the best signal of eventual contributions, indicating that updating commitments in response to new information may be critical for aligning collective efforts with a revised, more ambitious climate target.

As with all threshold experiments, questions of external validity arise notably in mapping our parameter values onto the real climate game. In our experiment, the values chosen for thresholds and endowments meant that groups had to contribute on average half of their endowments to reach the low threshold, and three-quarters of their endowments on average to reach the high threshold. Are these amounts commensurate with the average contributions required by countries to avoid crossing the 2°C and 1.5°C policy benchmarks? Similarly, is the range of 100 tokens in our uncertainty treatments equivalent to the threshold uncertainty countries are confronted with? It is extremely difficult to choose parameters that are numerically calibrated to the real climate game and for this reason, caution must be exercised in relating our results to the Paris Agreement policy benchmarks. Despite these limitations, a more confident and general conclusion from our work is that early warning signals indicating that a dangerous climate threshold lies closer than originally anticipated may increase the likelihood of avoiding crossing the original expected threshold, but may still be insufficient to prevent the revised one from being exceeded.

Taken together, our findings highlight both the promise and the limits of early warning signals as tools for fostering climate cooperation. Continued investment in methods to better quantify the probability of approaching dangerous thresholds is warranted, since even if early warnings cannot be detected in time to prevent a threshold from being crossed, they can still facilitate pre-emptive adaptation. But early warnings are not a silver bullet—strategic mechanisms for fostering cooperation remain essential, with early warnings serving at best as a complement rather than a substitute.

Methods

Experimental procedures

We recruited participants using the online participant crowdsourcing platform Prolific (N = 750; mean age = 39.74 years; SD = 12.02; range = 18–79; females = 351, males = 374, missing = 25). Participants received a participation fee of £6 and were informed that they would be playing a game with four other players in which they could earn a bonus payment. Ethical approval to conduct the experiment was granted by the Faculty of Science and Technology Research Ethics Committee at Lancaster University (FST-2025-5156-RECR-2).

The experiment employed a 6 (treatment: 100 vs. 50/150 vs. 150 vs. 100/200 vs. 135/165 vs. 145/155) × 10 (round: 1–10) mixed design: treatment was a between-groups factor, whereas round was a within-groups factor. Participants were tested in groups of five players. Groups were allocated at random to one of the six treatments, subject to the constraint of equal cell sizes (25 groups per treatment).

The experiment was executed using oTree[31], an open-source platform for running web-based interactive tasks. After reading an information sheet and providing electronic informed consent, participants were directed to a waiting page, where they remained until enough others arrived to form a 5-person group. Once grouped, players read the experimental instructions and completed a set of control questions (see Supplementary Experimental Instructions) to confirm their understanding of the rules of play. To preserve anonymity, each player was assigned a pseudonym (Ananke, Telesto, Despina, Japetus, or Kallisto). During the game, each player's decisions were communicated to the other players under their designated pseudonyms.

At the start of the game, each player received an endowment of 40 tokens. In each of ten rounds, players simultaneously and independently decided whether to contribute 0, 2, or 4 tokens from their endowment into an account for damage prevention. Players knew that if the total group contributions by the end of the game equalled or exceeded a threshold amount, then each player would get to keep the remaining balance of their endowment, issued as a bonus payment at a rate of 1 token = £0.50. However, if total contributions fell short of the threshold, then each player would lose 90% of their remaining endowment.

In treatments 100 and 150, the threshold was fixed at 100 or 150 tokens, respectively. In treatments 50/150, 135/165, and 145/155, the threshold was a random amount between 50 and 150 tokens. In treatment 100/200, it was a random amount between 100 and 200 tokens. In all uncertainty treatments, players knew that the exact threshold would be determined at the end of the game by drawing a whole number at random from the specified range, with each value being equally likely.

Before rounds 1 and 6, each player simultaneously and independently submitted two non-binding announcements. First, each player submitted a proposal regarding how many tokens the group should contribute in total over the ten rounds. They were then informed about their own proposal, the proposals of the other group members, and the average proposal. Players knew that the average group proposal would serve as the agreed collective target. Second, each player submitted a pledge regarding how many tokens they would personally contribute in total over the ten rounds. They were then informed about their own pledge, the pledges of the other group members, and the total pledge, alongside the group proposals to facilitate comparison.

At the end of each round, players were informed about their own contribution, their cumulative contribution, their most recent proposal and pledge, as well as the corresponding decisions made by the other group members. Summaries of the total round contributions, group total contributions, average group proposal, and total group pledges were also displayed. In this way, as the game progressed, players were able to gauge whether their group members were adhering to their pledges and whether the group contributions were consistent with achieving the agreed collective goal.

Before the second set of non-binding announcements, groups in treatments 135/165 and 145/155 were given an on-screen warning supplying them with updated information about the threshold range. Specifically, in treatment 135/165, groups were informed that the threshold was now a random amount between 135 and 165 tokens, whereas in treatment 145/155, they were informed that it was now a random amount between 145 and 155 tokens. In all other treatments, the fixed or uncertain threshold range remained as initially specified, and groups in these treatments received no additional information about the threshold. Instead, at the start of round 6, they proceeded directly to submit their second set of non-binding announcements.

At the end of the game, the threshold amount and the contents of the damage prevention account were communicated to the group. In the uncertainty treatments, the computer determined the exact threshold amount by drawing a random number from a uniform distribution either over the interval [50, 150] (treatment 50/150), [100, 200] (treatment 100/200), [135, 165] (treatment 135/165), or [145, 155] (treatment 145/155). After players had been informed of the outcome of the game and their payoff, they then completed a brief demographics questionnaire before being redirected to Prolific to receive their participation fee. Bonus payments were issued manually by the experimenter within approximately 12 hours of completing the session. More detailed methodological information can be accessed in the online resources (Supplementary Extended Experimental Methods and Supplementary Methodological Information).

Theoretical model

The imperfect information, repeated, and multiple player structure of the experimental game allows for a multiplicity of equilibria, and this complexity precludes a full equilibrium analysis. Therefore, we analyse the game under a set of simplifying assumptions and focus on two solutions: the internal cooperative equilibrium (which occurs when players maximise the sum of their individual expected net payoffs) and Nash equilibrium (which occurs occurs when each player maximises their own expected net payoff based on the expected contributions of others). This is possible because the game has a single pay-off period at the end and can therefore be partially analysed as an equivalent one-shot static game. A similar analysis of a one-shot game is provided by Barrett and Dannenberg[14].

We adopt the following simplifying assumptions:

1. All N players (countries) are identical and risk neutral.
2. In each round, players contribute q_{it}, which is the same in every round.
3. The total contribution (abatement) of N countries over T rounds i is: $Q_T = \sum_{i=1}^{N}\sum_{t=1}^{T} q_{it}$
4. The solutions presented are internal solutions that ignore the constraint on contributions per player, per round. Players are assumed never to contribute more in total than the upper bound of the uniform probability density function.

Table 3 provides a summary of the model parameters and variables. The structure of the model is as follows. The dangerous climate threshold is distributed $\tilde{Q} \sim f(\bar{Q},\varepsilon)$, where $f(\bar{Q},\varepsilon)$ is a uniform probability density function, $\bar{Q}$ is the mean of the distribution, and ε the dispersal around the mean. The probability of avoiding the threshold is given by the cumulative distribution:

$$P(\tilde{Q} \leq \sum_{i=1}^{N} Q_{iT}) = F(\bar{Q},\varepsilon,\sum_{i=1}^{N} Q_{iT})$$

The payoff depends upon whether total contributions exceed the realised threshold:

$$J_i(Q_{iT} \mid Q_{-iT},\tilde{Q}) = \begin{cases} \tau(X_{i0}-Q_{iT}) & Q_{iT}+Q_{-iT} < \tilde{Q} \\ X_{i0}-Q_{iT} & Q_{iT}+Q_{-iT} \geq \tilde{Q} \end{cases}$$

Relating this to the notation for a uniform distribution:

$$E[J_i(Q_{iT} \mid Q_{-iT})] = \begin{cases} \tau(X_{i0}-Q_{iT}) & Q_{iT}+Q_{-iT} < Q_{\min} \\ (X_{i0}-Q_{iT})\left[(1-\tau)F(\bar{Q},\varepsilon,Q_{iT}+Q_{-iT})+\tau\right] & Q_{iT}+Q_{-iT} \in [Q_{\min},Q_{\max}] \\ (X_{i0}-Q_{iT}) & Q_{iT}+Q_{-iT} > Q_{\max} \end{cases}$$

The payoff is restricted by contribution constraints in each round and in total. These limit the overall contribution and the rate at which the player can respond in a round. Thus, following Barrett and Dannenberg, $q_{it} \in [0, q_{max}]$. Over the planning horizon, the maximum contribution by a player is Tq_{max}. There is a further implicit constraint imposed in the game that the total contribution cannot exceed the initial endowment $Q_{iT} \in [0, X_{i0}]$.

The payoff for player i at the end of the planning horizon is the initial endowment X_{i0} less the total contributions up to the final round Q_{iT}. If the total contributions from all players are greater than or equal to the threshold $Q_{iT}+Q_{-iT} \geq \tilde{Q}$, the payoff is $X_{i0}-Q_{iT}$. If contributions are less than the threshold, the payoff is $\tau(X_{i0}-Q_{iT})$, where $0 \leq \tau \leq 1$ determines the penalty related to not achieving the realised threshold. The expected payoff for a player is given by:

$$E[J_i(Q_{it}|Q_{-it},Q)] = (X_{i0}-Q_{iT})F(\bar{Q},\varepsilon,Q_{it}+Q_{-it}) + \tau(X_{i0}-Q_{iT})(1-F(\bar{Q},\varepsilon,Q_{it}+Q_{-it})) \quad (1)$$

Eq. (1) can be rearranged so that the expected benefits term and expected costs are separated:

$$E[J_i(Q_{it}|Q_{-it},Q)] = X_{i0}(\tau+(1-\tau)F(\bar{Q},\varepsilon,Q_{iT}+Q_{-iT})) - Q_{iT}(\tau+(1-\tau)F(\bar{Q},\varepsilon,Q_{iT}+Q_{-iT})) \quad (2)$$

The cooperative solution is a special case in which Eq. (2) is maximised for a group of players that act as a single entity and the initial endowment is defined by $X_0 = X_{i0}N$:

$$E[J_i(Q_{iT}|Q)] = X_0(\tau+(1-\tau)F(\bar{Q},\varepsilon,Q_{iT})) - Q_{iT}(\tau+(1-\tau)F(\bar{Q},\varepsilon,Q_{iT}))$$

Table 3. Summary of model parameters and variables.

Description	Symbol	Value
Number of players	N	5
Initial endowment of each player	X_{i0}	40
Total group endowment	$X_0 = NX_{i0}$	200
Loss proportion if threshold not met	τ	0.1
Mean of the threshold distribution	$\bar{Q}$	Varies by treatment
Half-width of the threshold distribution	ε	Varies by treatment
Lower threshold bound	$Q_{\min} = \bar{Q} - \varepsilon$	Varies by treatment
Upper threshold bound	$Q_{\max} = \bar{Q} + \varepsilon$	Varies by treatment
Total contribution of player i	Q_{iT}	
Total contribution of all other players	Q_{-iT}	
Total group contribution	$Q_T = Q_{iT} + Q_{-iT}$	
Probability threshold is met	$F(\bar{Q}, \varepsilon, Q_T)$	
Payoff for player i	J_i	
Expected payoff for player i	$E[J_i]$	

Taking the derivatives with respect to Q_{iT}, and assuming an internal solution, yields the first-order marginal condition:

$$X_0(1-\tau)F(\bar{Q},\varepsilon,Q_{iT}) = (\tau+(1-\tau)F(\bar{Q},\varepsilon,Q_{iT})) + Q_{iT}(\tau+(1-\tau)F(\bar{Q},\varepsilon,Q_{iT}))$$

If we substitute in a cumulative uniform distribution with a lower bound $Q_{\min} = \bar{Q} - \varepsilon$ and an upper bound $Q_{\max} = \bar{Q} + \varepsilon$ the internal optimal cooperative solution is found by maximising the total payoff. The first-order conditions are:

$$Q_T^C = \frac{1}{2}\left(\frac{(Q_{\min} - \tau Q_{\max})}{(1-\tau)} + X_0\right) = \left(\frac{(X_0+\bar{Q})(1-\tau) - \varepsilon(1+\tau)}{2(1-\tau)}\right) \tag{3}$$

$$Q_T^C 2(1-\tau) - X_0(1-\tau) = \bar{Q}(1-\tau) - \varepsilon(1+\tau)$$

The share of the cooperative contribution per player is $Q_{iT}^C = Q_T^C/N$.

The Nash equilibrium contribution of a player is derived by differentiating Eq. (2) with respect to the individual's own contributions, assuming that all other players contribute equally. That is, $Q_{-iT} = (N-1)Q_{iT}$. The resulting Nash equilibrium contribution is:

$$Q_{iT}^N = \frac{(Q_{\min} - \tau Q_{\max})}{(1+N)(1-\tau)} + \frac{X_{i0}}{N(1+N)} = \frac{1}{1+N}\left(\frac{(X_{i0}+\bar{Q})(1-\tau) - \varepsilon(1+\tau)}{(1-\tau)}\right) \tag{4}$$

This result is multiplied by N to give the total contribution of all players:

$$Q_T^N = \frac{(1+N)}{N}\left(\frac{(X_{i0}+\bar{Q})(1-\tau) - \varepsilon(1+\tau)}{(1-\tau)}\right)$$

and rearranged to give:

$$Q_T^N(1-\tau) - X_{i0}(1-\tau) = \bar{Q}(1-\tau) - \varepsilon(1+\tau)$$

The righthand side of Eq. (4) can be equated with Eq. (3) to relate the Nash equilibrium to the cooperative equilibrium:

$$Q_T^C = \left(Q_T^N\frac{(1+N)}{2N} + \left(X_0 - \frac{X_0}{N}\right)\right)$$

The last step is to show that $Q_T^C - Q_T^N \geq 0$:

$$Q_T^C - Q_T^N = \left(Q_T^N\frac{(1+N)}{2N} + \left(X_0 - \frac{X_0}{N}\right)\right) - Q_T^N = -\frac{(-1+N)(Q_T^N - 2X_0)}{2N} \quad N = 1,2,3,\ldots,$$

If N=1 the Nash and cooperative contributions are identical. If $N \geq 2$, the first bracket is strictly positive and $(Q_T^N - 2X_0) < 0$ by assumption as the total endowment exceeds the total contribution. Thus, accounting for the initial sign, $Q_T^C - Q_T^N > 0$ for $N \geq 2$.

To generate the predictions shown in Table 2, we computed the Nash and cooperative equilibria for each treatment using the analytical solutions derived from our theoretical model. The disaggregated predictions for the first and second halves of the game were calculated as follows. In the low-threshold baseline treatments (100 and 50/150) and the high-threshold baseline treatments (150 and 100/200), predicted contributions were assumed to be evenly distributed over the first and second halves of the game. In the two early-warning treatments (135/165 and 145/155), we assumed players initially contributed in accordance with the low-threshold baseline uncertainty treatment (50/150), but revised their strategy in the second half of the game following the mid-game provision of more precise information about the threshold. Thus, predictions in these conditions reflect a shift in equilibrium behaviour following the arrival of the early warning signal.

Acknowledgements

This research was facilitated by a British Academy/Leverhulme Small Research Grant (SRG23\231641) awarded to MJH.

References

1. UNFCCC. *United Nations Framework Convention on Climate Change.* (1992).
2. Barrett, S. Why have climate negotiations proved so disappointing. *Sustain. Humanit. Sustain. Nature: Our Responsib. Pontif. Acad. Sci. Vatican City* 261–276 (2014).
3. UNFCCC. *United Nations Framework Convention on Climate Change. Copenhagen Accord.* (2009).
4. UNFCCC. *United Nations Framework Convention on Climate Change. Cancun Agreement.* (2010).
5. UNFCCC. *United Nations Framework Convention on Climate Change. Paris Agreement.* (2015).
6. Lenton, T. M. *et al.* Tipping elements in the earth's climate system. *Proc. Natl. Acad. Sci.* **105**, 1786–1793 (2008).
7. Richardson, K. *et al.* Earth beyond six of nine planetary boundaries. *Sci. Adv.* **9**, eadh2458 (2023).
8. Rockström, J. *et al.* A safe operating space for humanity. *Nature* **461**, 472–475 (2009).
9. Dannenberg, A. & Tavoni, A. Collective action in dangerous climate change games. In Dinar, A. *et al.* (eds.) *WSPC References of Natural Resources and Environmental Policy in the Era of Global Change, Vol. 4, Experimental Economics*, vol. 4 (World Scientific, 2017).
10. Hurlstone, M. J., Wang, S., Price, A., Leviston, Z. & Walker, I. Cooperation studies of catastrophe avoidance: implications for climate negotiations. *Clim. Chang.* **140**, 119–133 (2017).
11. Jacquet, J. Experimental insights: testing climate change cooperation in the lab. *Soc. Res. An Int. Q.* **82**, 637–651 (2015).
12. Barrett, S. Climate treaties and approaching catastrophes. *J. Environ. Econ. Manag.* **66**, 235–250 (2013).
13. Barrett, S. & Dannenberg, A. Climate negotiations under scientific uncertainty. *Proc. Natl. Acad. Sci.* **109**, 17372–17376 (2012).
14. Barrett, S. & Dannenberg, A. Sensitivity of collective action to uncertainty about climate tipping points. *Nat. Clim. Chang.* **4**, 36–39 (2014a).
15. Barrett, S. & Dannenberg, A. Negotiating to avoid 'gradual' versus 'dangerous' climate change: An experimental test of two prisoners' dilemmas. *Available at SSRN 2390561* (2014b).
16. Brown, T. C. & Kroll, S. Avoiding an uncertain catastrophe: climate change mitigation under risk and wealth heterogeneity. *Clim. Chang.* **141**, 155–166 (2017).
17. Dannenberg, A., Löschel, A., Paolacci, G., Reif, C. & Tavoni, A. On the provision of public goods with probabilistic and ambiguous thresholds. *Environ. Resour. Econ.* **61**, 365–383 (2015).
18. Hurlstone, M. J., White, B. & Newell, B. R. Threshold uncertainty, early warning signals and the prevention of dangerous climate change. *Royal Soc. Open Sci.* **12**, 240425 (2025).
19. Lenton, T. M. Early warning of climate tipping points. *Nat. Clim. Chang.* **1**, 201–209 (2011).
20. Lenton, T., Livina, V., Dakos, V., Van Nes, E. & Scheffer, M. Early warning of climate tipping points from critical slowing down: comparing methods to improve robustness. *Philos. Transactions Royal Soc. A: Math. Phys. Eng. Sci.* **370**, 1185–1204 (2012).

21. Lenton, T. M. Environmental tipping points. *Annu. Rev. Environ. Resour.* **38**, 1–29 (2013).
22. Scheffer, M. *et al.* Early-warning signals for critical transitions. *Nature* **461**, 53–59 (2009).
23. Scheffer, M. *et al.* Anticipating critical transitions. *Science* **338**, 344–348 (2012).
24. Tavoni, A., Dannenberg, A., Kallis, G. & Löschel, A. Inequality, communication, and the avoidance of disastrous climate change in a public goods game. *Proc. Natl. Acad. Sci.* **108**, 11825–11829 (2011).
25. Góis, A. R., Santos, F. P., Pacheco, J. M. & Santos, F. C. Reward and punishment in climate change dilemmas. *Sci. Reports* **9**, 16193 (2019).
26. Grimalda, G., Belianin, A., Hennig-Schmidt, H., Requate, T. & Ryzhkova, M. V. Sanctions and international interaction improve cooperation to avert climate change. *Proc. Royal Soc. B* **289**, 20212174 (2022).
27. Milinski, M., Semmann, D. & Krambeck, H.-J. Reputation helps solve the 'tragedy of the commons'. *Nature* **415**, 424–426 (2002).
28. Milinski, M., Semmann, D., Krambeck, H.-J. & Marotzke, J. Stabilizing the earth's climate is not a losing game: Supporting evidence from public goods experiments. *Proc. Natl. Acad. Sci.* **103**, 3994–3998 (2006).
29. Lenton, T. M. Tipping climate cooperation. *Nat. Clim. Chang.* **4**, 14–15 (2014).
30. Lenton, T. M. *et al.* The global tipping points report 2023 (2023).
31. Chen, D. L., Schonger, M. & Wickens, C. oTree: An open-source platform for laboratory, online, and field experiments. *J. Behav. Exp. Finance* **9**, 88–97 (2016).