# Collective Action to Avoid Catastrophe: When Countries Succeed, When They Fail, and Why

Scott Barrett
*Columbia University*

Special Issue Article

## Abstract

This article is concerned with situations in which avoiding a catastrophic outcome requires collective action. Using the logic of simple game theory, I identify circumstances that would cause rational players to act so as to guarantee that a catastrophe occurs, even when it is in their collective interests to avoid such an outcome – and when these players have the means at their disposal to ensure that such an outcome is avoided. I also identify circumstances that would cause rational players to act in concert to avert a catastrophe, and explain why other circumstances that might seem relevant to explaining these outcomes may be entirely inconsequential. Most of my discussion focuses on climate change. However, I also explain the relevance of the approach to three other areas: the millennium bug, drug resistance and nuclear arms control. In the final sections I discuss whether and how institutions might be designed to overcome, or at least reduce the likelihood of, catastrophic collective-action failures at the international level.

## Policy Implications

- The ability of countries to organize to avoid catastrophes depends critically on uncertainty about the threshold, or tipping point, for catastrophic change.
- When this uncertainty is small, avoiding catastrophe requires coordination – something countries are very good at doing.
- When this uncertainty is large, collective action requires enforcement of a cooperative agreement – something countries are very bad at doing.
- Enforcement can be enhanced by countries ceding some sovereignty – and yet, historically, countries have been unwilling to do this without having first experienced a catastrophic outcome.
- In some cases it may be possible to devise strategies that transform a collective-action problem into a coordination game.

Some catastrophes come from out of the blue, and there is not much that humans can do about them except prepare for their eventuality, devise early warnings where feasible, and limit the harm caused after catastrophe has struck. The Boxing Day tsunami that hit the Indian Ocean in 2004 is an example of this kind of catastrophe. An even more dramatic example is the super-eruption of Mount Toba about 74,000 years ago – an event that may have triggered a 'volcanic winter' akin to a 'nuclear winter', with devastating consequences, possibly including a population collapse leading to a 'genetic bottleneck'.

My concern in this article is with a different kind of event: catastrophes that are either caused by humans or that can be prevented by human actions. An example is 'catastrophic' climate change, such as the melting of the Greenland ice sheet, caused by a failure to limit greenhouse gas emissions. Events like these often have thresholds. They occur only if the threshold is crossed.

Of particular importance is the way in which human actions determine *whether* a critical threshold is crossed. A single country or 'coalition of the willing' could keep global temperatures on the safe side of a critical threshold by using geoengineering – that is, by throwing particles into the stratosphere to deflect sunlight (Keith, 2000). However, not every catastrophic event associated with global warming can be averted by this means. Moreover, an attempt by one country or a group of like-minded countries to interfere in the climate in this way may only provoke counter-reactions by other states. It is not obvious that geoengineering is a true 'solution' to catastrophic climate change (Barrett et al., 2014).

Global Policy (2016) 7:Suppl.1 doi: 10.1111/1758-5899.12324

Limiting global temperature by reducing greenhouse-gas emissions, the root cause of climate change, is a safer bet. However, doing this requires the efforts of a large number of countries. It requires global collective action (Sandler, 1997, 2004; Barrett, 2003, 2007) – a kind of global remedy for the 'tragedy of the commons' memorably told by Hardin (1968) and compellingly analyzed at the local level by Ostrom (1990).

A key feature of many catastrophes is uncertainty. If temperatures rise too high, the Greenland ice sheet will melt. This much is known. What is not known is the precise change in mean global temperature that would cause the Greenland ice sheet to melt (1–2°C relative to recent temperature, according to Lenton et al. (2008); see also Kriegler et al. (2009)). Also uncertain is the increase in atmospheric concentrations that would give rise to this change in temperature (a phenomenon known as 'climate sensitivity'), and the increase in global emissions that would push atmospheric concentrations over this critical concentration level (an amount that is uncertain due to unpredictability in the carbon cycle). I call this *threshold uncertainty*.

Another uncertainty concerns the *consequences* of an event caused by crossing such a threshold. Whether melting of the Greenland ice sheet turns out to be truly 'catastrophic', for example, depends on the extent of sea-level rise attributable to this event (2–7 meters, according to Lenton et al. (2008)), the 'background' level of sea-level rise that will occur regardless, the speed at which the Greenland ice sheet melts (from 300 to more than 1,000 years, according to Lenton et al. (2008)), the human actions taken to limit the physical impacts associated with this sea-level rise (adaptation), and the values that people put on these impacts. I call this *impact uncertainty*.

In this article I develop a way to think about global collective action to avert a catastrophe. My main conclusion is that, due to threshold uncertainty but not impact uncertainty, collective action to avert a catastrophe is likely to fail, even when the world would be much better off if catastrophe were averted. Although the problem appears to be behavioral, I would argue that it is fundamentally institutional. If you think of institutions as 'the rules of the game in a society or, more formally . . . the humanly devised constraints that shape human interaction' (North, 1990, p. 3), the central problem is that our institutions haven't equipped us to avert every catastrophe.

## Avoiding catastrophe

In an interesting and provocative article, Milinski et al. (2008) report the results of an experiment intended to simulate the 'collective-risk social dilemma' in preventing 'dangerous climate change'. There are six players. Each is given €40. The game is played in ten periods. In each period, every player must choose to contribute €0, €2 or €4. If, at the end of the game, at least €120 has been contributed, dangerous climate change is averted with certainty, and each player gets a payoff equal to the amount of money he or she has left (there are no refunds in this game). If less than €120 has been contributed, each player loses all the money he or she has left with probability of 0.9. In their experiment, Milinski et al. (2008) played the game with ten groups of students. In these experiments, only half of the groups averted 'catastrophe'.

How could this happen? The problem here is not with people, but with the rules of the game.

In this game, there are two symmetric pure strategy Nash equilibria. In one, every player contributes €0 every period. This gives each player an expected payoff of €4. In the other equilibrium, every player contributes €2 every period. This gives every player a certain payoff of €20. (Of course, many asymmetric pure strategy equilibria also exist in which different players contribute different amounts, possibly in different periods.[1] ) Of the two equilibria, only the latter one is efficient – the collective best outcome.[2]

Why do the players not consistently obtain this better outcome? The main reason is that, as the game is constructed, the players are not allowed to communicate. Whether players can or cannot communicate is one of the rules of the game. Clearly, if there is one thing climate negotiators do, it is to communicate.

Communication matters in this game because the fear of crossing a catastrophic threshold gives the players a strong incentive to coordinate their actions. Without communication, each player is a little unsure of what to do, because each is unsure about the intentions of the other players. It is obvious from the results of this experiment that the prohibition on communication is a problem. The groups that failed to avoid the threshold with certainty contributed €113 on average, just €7 short of the full cooperative outcome. This average outcome gave each player an expected payoff of €2.12, which is lower than each player would have got had every player contributed nothing and substantially less than each would have got had every player contributed just a little bit more. Had the players been able to start again, they would certainly have chosen differently; had they been able to communicate from the start, they would almost certainly have chosen differently. Tavoni et al. (2011) demonstrated this in a follow-up experiment. When the players in this game were allowed to communicate, they showed, groups were much more likely to avoid catastrophe.

The institution of a treaty could help even more. The usual way of modeling an international environmental agreement is in three stages (Barrett, 2003). In stage 1, countries choose independently whether to be a party or nonparty. In stage 2, parties choose their actions (in this case, contributions) so as to maximize their collective payoff. Finally, in stage 3, nonparties choose their actions with the aim of maximizing their individual payoffs. A treaty is self-enforcing if, given the treaty and participation level, nonparticipants do not want to change their behavior; if, given the participation level, parties to the treaty do not want to change the obligations expressed in the treaty; and if, given the participation decisions of other countries, each country does not want to change its decision of whether to be a party or nonparty to the treaty.

Applying this notion of a self-enforcing treaty to the Milinski et al. (2008) game, it will help to simplify matters. Assume that contributions are made in a single period, and that each player can contribute any amount of money up to his or her endowment. If participation were full, the treaty would then tell each country to contribute €20, netting each country a payoff of €20. Were a country to drop out of this agreement, the remaining parties would change their contributions. They would reason that if they contributed an amount $Y$ in total, then (taking this contribution as given) the nonparty would contribute an amount $Z = €120 - Y$ for $€120 \geq Y \geq €84$ and $Z = €0$ for $€84 > Y > €120$. Knowing this, the five remaining signatories could do no better than to contribute $Y = €84$ collectively – that is, €16.80 each. The remaining cooperators would lower their contribution level to punish the one country for withdrawing. This would net each of the five parties €23.20, whereas the sole nonsignatory would get just €4. Recall that, were this country not to withdraw, it would get a payoff of €20. Obviously, with the treaty written in this way, no country would have an incentive to withdraw, starting from a situation in which participation was full. The treaty comprising six signatories, each of which contributed €20, would thus be self-enforcing.

The treaty just described serves as a *coordination device*. Communication steers the parties towards the better outcome, but a treaty provides complete assurance as to the likelihood of achieving this outcome. Treaties are very good at coordinating behavior. What treaties are not good at is enforcement. In this case, however, enforcement is provided compliments of Mother Nature. Should the parties not contribute enough in total, there is a very high chance that they will all be punished by triggering a catastrophe. Thanks to the prospect of catastrophe, avoiding catastrophe is very highly probable.

Is the game just described the one that countries are now playing? Unfortunately, as I shall now explain, the true game of climate change is very different from this.

## Uncertainty reconsidered

In the Milinski et al. (2008) game, the threshold is certain. It is €120. What is uncertain is the impact of crossing the threshold. If the players contribute less than €120 in total, they have a 90 per cent chance of losing everything that they contributed and a 10 per cent chance of losing nothing.

If you think of contributions as being 'abatement of greenhouse gases', then this money will be lost as soon as it is spent, although it may help by reducing 'gradual climate change' even if it fails to avert 'catastrophe'. The impact of crossing a threshold will be independent of this expenditure. For example, if you think of the threshold as triggering the disintegration of the West Antarctic ice sheet, the impact of crossing the threshold will be the costs of rapid (in geological terms) sea-level rise. Presumably, this cost will be much greater than the amount of money spent to avoid the threshold with certainty (again, €120). Otherwise, it would not make any sense to avoid 'catastrophe'.

In a theoretical article, I have modeled the problem in a different way (Barrett, 2013). Suppose that the threshold is defined in terms of abatement from 'business as usual' emissions. Supplying enough abatement is thus equivalent to limiting cumulative emissions so as to stay on the 'safe' side of a 'dangerous' level of atmospheric concentrations of greenhouse gases. This amount is determined by nature, but the players in this game may not know the threshold – just as, in the real climate-change game, tipping points for major geophysical features are uncertain (Lenton, 2008). In my model abatement reduces 'gradual' climate change, but is only able to avoid 'catastrophic' climate change if this (possibly unknown) threshold is avoided. The impact of crossing the threshold is a value determined independently of the amounts spent to avoid the threshold. This impact may also be uncertain.

In this model, whether crossing a threshold is truly 'catastrophic' depends on the parameter values, meaning the situation at hand. To be truly 'catastrophic', the costs of avoiding the threshold must be low relative to the impacts of crossing the threshold. Not all thresholds are worth avoiding. Some are very much worth avoiding.

Consider first the situation in which both the threshold and the impact are certain. Then, when the economics of avoiding the threshold are barely favorable, avoiding catastrophe will be a prisoners' dilemma game. When the economics of avoiding the threshold are strongly favorable, however, avoiding catastrophe will be a coordination game. A treaty may help to overcome the prisoners' dilemma under very limited circumstances, but even in these situations a treaty will improve welfare very little (by definition, the economics of avoiding catastrophe are barely favorable when catastrophe avoidance is a prisoners' dilemma). By contrast, when avoiding catastrophe is a coordination game, a treaty can be relied upon to ensure that the threshold is avoided.

What happens when there is uncertainty? In particular, what happens when the threshold and impact are uncertain but have expected values that are identical to their true values in the certainty case? Consider first the full cooperative outcome. Compared to the certainty case, uncertainty about the impact of crossing the threshold makes no difference; the total abatement level that makes the world as well off as possible is unchanged. This is because optimization is assumed to depend only on the expected value of the impact (that is, preferences are assumed to be risk-neutral). By contrast, uncertainty about the threshold matters a lot. It may pay to reduce the chance of catastrophe dramatically, increasing the level of abatement substantially relative to the certainty case. Indeed, depending on the probability-density function for the threshold, it may pay to reduce emissions by so much that the critical threshold is *certain* to be avoided.

This perspective is useful for understanding the controversial 'precautionary principle'. Even without any preference for avoiding risk, it may pay the world to take stronger measures to avoid a risk that is more uncertain. Uncertainty about the impact plays no role here. It is only uncertainty about the threshold that matters.

What is the effect of uncertainty on collective action? The model makes a very strong prediction. Uncertainty about the impact should have no effect on collective action. By contrast, uncertainty about the threshold should be critical. Indeed, while uncertainty about the threshold can make it collectively desirable for the world to abate more than in the certainty case, collective action may cause countries to abate much less. The failure of collective action due to threshold uncertainty can cause countries to act as if the risk of catastrophe can be ignored, even when this risk is very great.

What is the intuition behind these results? Countries can do nothing to reduce uncertainty about the impact, and so their behavior is unaffected by this uncertainty. To be clear, the impact of crossing the threshold matters very much, but decision-making is affected only by the expected value of the impact, not its uncertainty. By contrast, countries can reduce the chance of crossing the threshold. In the full cooperative outcome, this uncertainty makes countries want to abate more compared to the certainty case. However, when uncertainty about the threshold is 'large', abatement is a prisoners' dilemma. Under these circumstances, countries have a strong collective incentive to abate more (compared to the certainty case), but a strong individual incentive to abate less. When the threshold is certain, if every other country plays its part in avoiding the threshold, each country wants to play its part. The reason is that, should a particular country not play its part, the threshold is certain to be crossed. By contrast, when the threshold is uncertain, even if every other country plays its part in reducing the risk of crossing the threshold, each country has an incentive to scale back on its abatement. By doing so, a country reduces its abatement costs significantly but increases the chance of catastrophe only slightly. Tragically, when every country faces this same incentive and behaves in this same way abatement drops substantially – making catastrophe nearly certain.

## An experimental test

Would real people, playing for real stakes, behave as this theory predicts? To find out, Astrid Dannenberg and I tested the theory in a computer laboratory at the University of Magdeburg, Germany, using students recruited from the general student population (Barrett and Dannenberg, 2012). Each student was randomly assigned to a group of ten students, and each experimental 'treatment' was conducted for ten groups. As there were four treatments, a total of 400 students participated in the experiment. Our experiment strongly confirmed the theoretical predictions of the model.

Here I provide an informal description of the game played in the lab. (For details about how the actual experiment was conducted and its results, see Barrett and Dannenberg (2012).) Every player was given ten black and ten red poker chips. Players received €0.10 for every black chip they kept and €1 for every red chip they kept. They also got €0.05 for every chip handed in by the group, with no distinction being made as to the color. On top of this, every player lost €15 if the total number of chips contributed was less than 150. In the *certainty* treatment, the threshold was 150 and the impact of crossing it was €15.

We also conducted three other treatments; see Table 1 for a comparison. With *impact uncertainty*, the impact of falling short of the threshold was not €15 precisely but a value chosen randomly between €10 and €20. With *threshold uncertainty*, the threshold was not 150 but a real number chosen randomly between 100 and 200. Finally, in the *impact and threshold uncertainty* treatment, both the impact and the threshold were uncertain. Notice that in the uncertainty treatments, each parameter's expected value was always equal to its actual value under certainty.

There is one final point to note: the players were allowed to communicate before making their choices. In particular, each person was allowed to propose how much he or she thought the group should contribute. Each person was also allowed to make a pledge for how much he or she intended to contribute individually. These proposals and pledges were nonbinding.

**Table 1.** Collective action to avoid catastrophe

| Treatment | Threshold | Impact | Prediction | Result |
|---|---|---|---|---|
| Certainty | 150 | €15 | Avoid catastrophe | Catastrophe avoided 8 out of 10 times |
| Impact uncertainty | 150 | €10–€20 | Avoid catastrophe | Catastrophe avoided 10 out of 10 times |
| Threshold uncertainty | 100–200 | €15 | Catastrophe | Catastrophe occurred for certain 9 out of 10 times and with probability 0.93 the other time |
| Impact and threshold uncertainty | 100–200 | €10–€20 | Catastrophe | Catastrophe occurred for certain 7 out of 10 times and with probability 0.80–0.91 the other times |

*Source: Barrett and Dannenberg (2012).*

As indicated in the table, the behavior observed in the laboratory strongly confirms the predictions of the theory. The theory predicts that behavior should be indistinguishable between the certainty and impact uncertainty treatments – in these cases, the players should be able to coordinate to avoid catastrophe – and our results are consistent with this prediction. The theory predicts that behavior should be indistinguishable between the threshold uncertainty and impact and threshold uncertainty treatments, and our results are consistent with this prediction. Finally, the theory predicts that behavior should be very different between these pairs of treatments, and this prediction is also confirmed by how people play in the lab. People can coordinate to avoid catastrophe when the threshold is certain. They cannot coordinate to avoid catastrophe when the threshold is uncertain. Uncertainty about the impact of crossing a threshold has no effect on collective action.

Actually our results are even stronger than suggested by Table 1. It turns out that in each of the cases in which catastrophe occurred in the certainty treatment, the reason was that a single individual contributed substantially less than he or she pledged. Overall, for both of these treatments, 98 per cent of the players contributed at least as much as they pledged. By contrast, in the two treatments with threshold uncertainty, the vast majority of players contributed less than they pledged. Finally, while our results indicate that a few groups were able to lower the probability of catastrophe slightly, in every one of these cases catastrophe occurred ex post once 'nature' chose the actual value for the threshold. (In our experiment, this value was chosen by asking one of the participants to activate a computerized 'spinning wheel' programmed to choose a random value between the end points of the uniform distribution corresponding to the threshold parameter.)

## Reducing uncertainty: early warnings

These results contrast certainty with uncertainty. As noted earlier, however, collective action to avoid a dangerous threshold is a prisoners' dilemma only if the threshold is sufficiently uncertain. If uncertainty is small enough, collective action to avoid catastrophe will be a coordination game.

While thresholds are inherently uncertain in most cases, there is some evidence that uncertainty shrinks as a threshold is approached. In short, it may be possible to detect 'early warning signals' of an approaching catastrophe.

To test the sensitivity of collective action to uncertainty about the threshold for a regime shift, Astrid Dannenberg and I conducted another experiment (Barrett and Dannenberg, 2014a). Knowing now that impact uncertainty does not affect behavior, in this experiment we focused exclusively on threshold uncertainty. Using the theory developed in Barrett (2013), we identified a 'dividing line' value for threshold uncertainty. Inside the dividing line, collective action is a coordination game, and theory predicts that players will be able to avoid catastrophe. Outside the dividing line, collective action is a prisoners' dilemma, and theory predicts that efforts to avoid catastrophe will fail.

The results of our experiment are summarized in Table 2. There are five treatments, each with a different level of threshold uncertainty. These range from zero uncertainty (a threshold of 150) to high uncertainty (a threshold of 100–200). These two treatments were taken from the experiment discussed in the previous section. The other treatments were new. One of these (145–155) represents uncertainty just inside the dividing line, which is predicted to 'tip' group behavior towards coordination (the predicted dividing line being 142–158). Another treatment (140–160) is just outside the dividing line. Even though this range of values is very close to those in the 145–155 treatment, the theory predicts that, in this case, cooperation should collapse. The same result is expected for the final treatment (135–165).

As shown in Table 2, these predictions are strongly supported by the experimental results. The nonlinear nature of the behavioral response to uncertainty suggests that a threshold must be known with near certainty in order for collective action to shift from a prisoners' dilemma to a coordination regime.[3]

**Table 2.** Sensitivity of collective action to uncertainty about the threshold

| Threshold treatment | Theoretical prediction | Experimental result |
|---|---|---|
| 150 | Avoid catastrophe | Catastrophe avoided 8 out of 10 times |
| 145–155 | Avoid catastrophe | Catastrophe avoided 4 out of 10 times for certain and with probability ranging 0.30–0.80 the other times |
| 142–158 | Dividing line | |
| 140–160 | Catastrophe | Catastrophe occurred for certain 10 out of 10 times |
| 135–165 | Catastrophe | Catastrophe occurred for certain 10 out of 10 times |
| 100–200 | Catastrophe | Catastrophe occurred for certain 9 out of 10 times and with probability 0.93 the other time |

*Source: Barrett and Dannenberg (2014a).*

For many problems, this sensitivity is probably too great to support the kind of behavioral change needed to avert disaster. For example, in the case of climate change, early warning signals may fail completely or be prone to false positives and false negatives (Lenton, 2011). Even when they work, they may not reduce uncertainty by enough to bring about the behavioral change needed to avert disaster. Another problem, of course, is that by the time early warning signals are detected, there may be too little time to take the actions needed to avoid the threshold. There may be enough time to adapt to climate change or to deploy geoengineering, but it takes decades for reductions in greenhouse-gas emissions or even $CO_2$ removal from the atmosphere to affect temperature.

## Relevance to the climate-change negotiations

The results reported here are entirely consistent with how the climate negotiations have developed over the last 25 years. Beginning with the United Nations Framework Convention on Climate Change (UNFCCC), adopted in 1992, negotiators framed their problem as needing to avoid 'dangerous' climate change. In the 2015 Paris Agreement, they quantified this threshold, establishing the goal of 'holding the increase in the global average temperature to well below 2°C above pre-industrial levels and to pursue efforts to limit the temperature increase to 1.5°C above pre-industrial levels, recognizing that this would significantly reduce the risks and impacts of climate change'. Under the Paris Agreement each country has also made a pledge, known as an intended nationally determined contribution (INDC), for what it intends to do up to 2025 or 2030 to ensure that this threshold is avoided. An analysis of these pledges by the UNFCCC Secretariat (2015) finds that 'much greater emission reductions effort than those associated with the INDCs will be required in the period after 2025 and 2030 to hold the temperature rise below 2°C above pre-industrial levels' (p. 11). To be more precise, the Secretariat's analysis predicts that, even assuming that the INDCs are met, global emissions will increase through 2030. To have a greater than 50 per cent chance avoiding the threshold, this trend would have to be reversed starting in 2030. And this is assuming that all the INDCs will be met. It will take many years before we will know what countries end up doing, but if the theory and experiments summarized here are any guide, we should not be surprised if many countries fall short of meeting their pledges. It thus seems very likely that the 2°C threshold for averting dangerous climate change will be exceeded.

## Additional case studies

Up to this point, I have discussed this problem only in the context of climate change. However, the theory should apply more generally. I will now consider three more applications.

### The year 2000 problem

The Y2K bug is an example of an approaching catastrophe with a certain threshold (the date 1 January 2000). Of course, the Y2K bug turned out not to be the catastrophe that some people had expected it to be, but the reasons for this are unclear. Catastrophe might not have occurred because of the actions that were taken to avoid it. Alternatively, the threat of calamity may have been overblown all along. The theory presented here is consistent with both interpretations. Although the threshold was certain, the impacts were very uncertain. In expected value terms the impacts were significant. In a worst case, the impacts were scary. In a best case, they were negligible. What is interesting and perhaps unique about this case is that the threshold was certain. Applying the theory outlined previously, countries should have coordinated their actions to avoid catastrophe.

When the Y2K alarm was raised, countries mobilized a significant coordination effort. They established the International Y2K Cooperation Center under the auspices of the United Nations and turned to the World Bank for funding. As explained in a press release accompanying its final report, the Center's role was 'to *coordinate* [my emphasis] efforts to update computer and automated control systems around the world to transition smoothly to the year 2000'.[4] In its final report, the Center explained the reasons for its success:

> Unprecedented international cooperation contributed to the successful outcome. Two attributes of the Y2K problem helped make that cooperation possible. First, Y2K threatened every nation, providing the incentive to share best practices and reduce the total costs of fixing the problem. The unmovable deadline of 1 January 2000 gave gravity to this menace. Second, it would do a country limited good to solve its own problems if a neighbor on whom it depended for critical services or supplies was unable to function because of Y2K failures. The interdependency among nations created interest in providing mutual assistance so that all could succeed.[5]

Probably the greatest fear about Y2K was the possibility that nuclear missiles might be launched by accident, or that a computer might falsely detect an approaching missile, triggering a nuclear counterattack. To avert such a danger, the US and Russia not only chose to be in close contact about the things each country was doing to protect against the Y2K bug; they also established a joint center to detect false warnings of missile attacks. The nature of this coordination is virtually without precedent:

> At Peterson's Air Force Base in Colorado, Russian and US military personnel sat side by side as part of a pioneering missile watch. The specialists shared workstations beginning on December 27, 1999, and kept vigil in shifts of 20 personnel until mid-January 2000. Throughout the watch, the military officers were in telephone contact with command centers in both the US and Russia (Smith, 2000, pp. 194–195).

With this arrangement, not only did the US and Russia both know what was going on, but all of this information was in full view. That is, each party knew what the other party knew about what was going on; each party knew that the other party knew what it knew about what was going on; and so on. The arrangement allowed the two sides to coordinate perfectly, ensuring that the Y2K threshold could not trigger a nuclear war.

## Drug resistance

Drug-resistant microbes have high evolutionary fitness in the presence of a drug. However, the acquisition of resistance may entail a fitness cost that is harmful to the pathogen when the drug is removed.[6] This means that if the use of the drug is low the drug-sensitive strains will tend to win out, allowing the drug to remain effective; if use of the drug is high the resistant strains may thrive, making the drug ineffective. An implication is that resistance can be controlled by the extent of drug use in a population. To prevent resistance from emerging and spreading, use of the drug must be kept below some critical threshold. However, if there is only one drug available, this means not treating people who are sick in order to allow more sick people to be treated in the future – a difficult decision. When there are multiple drugs that work, other strategies can be tried. These include cycling of different drugs (never using any single drug for very long), using different drugs at the same time (perhaps prescribing one for adults and another for children) or using combinations of drugs (it is harder for a pathogen to acquire resistance to multiple drugs at the same time).

*Plasmodium falciparum*, the most harmful form of malaria, has developed resistance to most antimalarial treatments, including chloroquine, sulfadoxine-pyrimethamine and mefloquine. Today artemisinin therapies are the most effective treatment available, but their future effectiveness is in danger. When administered as monotherapies, these drugs are prone to resistance. Artemisinin-based combination therapies make resistance less likely, partly because the partner drug has a longer half-life and so clears more of the parasites from the body. But for a variety of reasons, including cost, people often use single-drug treatments when they are available. Fearing a catastrophe, should resistance emerge and spread, the World Health Assembly adopted a resolution in May 2007 urging states to 'cease progressively the provision in both the public and private sectors of oral artemisinin monotherapies'.

Unfortunately, soon after this resolution was passed, artemisinin resistance was observed in the area bordering Cambodia and Thailand. (Since then resistance has been observed elsewhere, including on the Thailand–Myanmar border.) There is a fear that these resistant strains could spread to Africa, where *P. falciparum* kills large numbers of children. As Margaret Chan, director-general of the World Health Organization, put it: 'it is no exaggeration for me to say that the consequences of wide-spread resistance to artemisinins would be catastrophic'.[7]

If avoiding resistance were a coordination game, unanimous agreement by the World Health Assembly should have sufficed to ensure decisive action. One problem may well have been uncertainty about the threshold. However, even after resistance emerged, the response was weak. A number of countries, including Cambodia, have imposed and even enforced a ban on the proscribed drugs. But many countries have failed to take official action.[8] And many of those that have taken official action have failed to enforce their regulations. More than two dozen companies continue to produce these drugs.[9]

Why the inadequate response? One reason is that the World Health Organization lacks the authority to enforce a worldwide ban. But another is that many countries have weak or fragile domestic-governance arrangements. These countries do not possess the ability to enforce domestic laws prohibiting the distribution of single-drug medicines.[10] A key assumption of the theory is that the players in a game have the wherewithal to act, but in the real world this assumption cannot always be relied upon.

### Nuclear weapons

Of all the global-scale catastrophes that might be contemplated, none is perhaps more dreaded, or more likely, than nuclear war.

One concern is that nuclear war on a very large scale could trigger a climate catastrophe, possibly leading to the extinction of humanity. Carl Sagan identified 'a crude threshold, very roughly around 500 to 2,000 warheads', and recommended limiting arsenals to 'a value below the low end of the plausible range' (Sagan, 1983, p. 285). According to Sagan:

> National or global inventories above this rough threshold move the world arsenals into a region that might be called the 'Doomsday Zone'. If the world arsenals were well below this rough threshold, no concatenation of computer malfunction, carelessness, unauthorized acts, communications failure, miscalculation and madness in high office could unleash the nuclear winter. When global arsenals are above the threshold, such a catastrophe is at least possible. The further above threshold we are, the more likely it is that a major exchange would trigger the climatic catastrophe (Sagan, 1983, pp. 285–286).

Although the impact of all-out war is uncertain, and may turn out to be less than apocalyptic (Thompson and Schneider, 1986), the theory sketched previously suggests that threshold uncertainty is the greater problem. However, the incentive problems with avoiding catastrophe shrink considerably as the number of players is reduced. Today there are nine countries with nuclear weapons, and only two of these (Russia and the US) hold stocks above the Sagan thresholds.[11] Moreover, each of these countries has a powerful self-interest motive for avoiding nuclear war and for reducing their stockpiles jointly.[12]

In 1974, the US and USSR negotiated the Threshold Test Ban Treaty, an agreement that established a nuclear testing 'threshold' by prohibiting tests having a yield exceeding 150 kilotons (equivalent to 150,000 tons of TNT), 'the fractional-megaton range'. This limit was important because only larger weapons would be used in a first strike, and if a state could not be sure that its large weapons would work then it would be less willing to launch an attack in the first place.

The threshold specified in this agreement is certain, but verifying compliance with the threshold is uncertain. One problem is that underground tests can trigger tectonic release, making it difficult to know whether a bomb has been tested. Another problem is that the amplitude of a measured seismic wave is only loosely related to the yield of the weapon that is tested. Although these uncertainties are limited, and did not inhibit entry into force of this bilateral agreement, they have been invoked to block ratification of the Comprehensive Test Ban Treaty.

## Institutional correctives

I have tried to show that avoiding catastrophe presents an important problem for global collective action. Too little is done to avert a catastrophe. To change things, our international institutions must become better at overcoming the prisoners' dilemma of avoiding catastrophe (Barrett, 2006).

Richard Posner has proposed a radical remedy. He says that, to address problems like global climate change, we need an 'international environmental protection agency' with autonomous enforcement powers. He admits that such an agency 'would involve a significant surrender of sovereign powers on the part of the nations of the world – which is probably why there is no such agency'. But he believes that 'there may be no feasible alternative means of curbing highly destructive global negative externalities' (Posner, 2004, p. 216).

Under what circumstances might countries be willing to cede a measure of their sovereignty? A simple and possibly simplistic reading of history suggests that countries are unlikely to accept dramatic institutional changes before a crisis has hit. It was only after the Second World War that countries adopted the United Nations Charter; only after the 2004 SARS outbreak that members of the World Health Organization revised the International Health Regulations; and only after the euro crisis that the euro members agreed to the Fiscal Compact. Each one of these institutional reforms or changes limited the sovereignty of member states, helping them to act to avert a future crisis. However, in every case, countries were unwilling to cede any of their sovereignty before they had experienced a catastrophic outcome.

## Strategic approaches

The thresholds considered thus far were all given exogenously. Might a threshold be determined strategically? To avert a catastrophic nuclear war, Herman Kahn proposed the construction of a doomsday machine:

> A device whose function is to destroy the world. This device is protected from enemy action (perhaps by being situated thousands of feet underground) and then connected to a computer, in turn connected to thousands of sensory devices all over the US. The computer would be programmed so that if, say, five nuclear bombs exploded over the US, the device would be triggered and the world destroyed. Barring such problems as coding errors (an important technical consideration), this machine would seem to be the 'ideal' [deterrent]. If Khrushchev ordered an attack, both Khrushchev and the Soviet population would be automatically and efficiently annihilated (Kahn, 1961, p. 107).

A climate doomsday machine would connect all the world's nuclear bombs to a computer, which in turn would be linked to a sensor atop Mona Loa in Hawaii. This is where readings are taken of atmospheric concentrations of greenhouse gases. Today, the concentration level is about 400 parts per million by volume (ppmv). The computer could be programmed to destroy the world should this level top, say, 500 ppmv. With the trigger for catastrophe being certain, theory and experimental evidence strongly suggest that this device would give the world all the encouragement needed to stay within 500. The world would get around natural uncertainty about the threshold by creating artificial certainty.

Of course, I'm not seriously proposing that the world do this.[13] I only offer the example as a thought experiment. The climate doomsday machine is a pure strategic device. Its sole purpose would be to change the incentives countries have to rein in their emissions and save the world from dangerous climate change. It works by transforming the prisoners' dilemma into a coordination game.

Are there acceptable ways in which negotiators could turn a problem like climate change into a coordination game? The Montreal Protocol on Substances that Deplete the Ozone Layer is one of the most successful international agreements ever adopted. The secret of its success lies in its ability to enforce participation by threatening to restrict trade in the controlled substances (mainly chlorofluorocarbons (CFCs)) between parties and nonparties (Barrett, 2003). Like the doomsday machine, the trade restrictions in the Montreal Protocol are a strategic device. Their purpose is to make phasing out CFCs into a coordination game.

How do the trade restrictions work? Imagine that very few countries belonged to an agreement to limit CFCs. Would your country wish to join the agreement? With the trade restrictions in place, the answer is almost certainly no. By joining, your country would not only forfeit the advantage of free riding, it would also – thanks to the trade restriction – suffer a loss in the gains from trade. Imagine, however, that nearly every country participated in such an agreement. Would your country wish to join then? By joining, your country would still forfeit the advantage of free riding but now, by joining, your country would be able to capture all of the gains from trading in CFCs with the rest of the world. Joining would be in your country's interests provided the trade gains exceeded the loss from foregoing free riding.

Notice that the participation game is characterized by tipping. If each country expects few others to participate, none participates. If each country expects almost all others to participate, all countries participate. Somewhere in between these outcomes lies a 'tipping point' for participation. To coordinate participation, the treaty must simply set the minimum participation level for entry into force equal to a value greater than the tipping point (Barrett, 2003).

The trade restrictions in the Montreal Protocol are an acceptable alternative to the doomsday machine. Like the doomsday machine, the trade restrictions transform the prisoners' dilemma into a coordination game.

Can the climate problem be addressed in this same way? It turns out that the most effective climate treaty so far was not the Kyoto Protocol but the Montreal Protocol, which was never intended to limit climate change. Calculations by Velders et al. (2007) show that, by phasing out the ozone-depleting substances that double as greenhouse gases, the Montreal Protocol achieved four times as much as the Kyoto Protocol aimed to achieve in terms of limiting climate change. Kyoto could not be enforced, but the Montreal Protocol has been enforced thanks to trade restrictions. In November 2015, a decision was made to negotiate an amendment to the Montreal Protocol to phase down hyrdrofluorocarbons (HFCs), a chemical that is harmless to the ozone layer but a potent greenhouse gas – a gas that the Kyoto Protocol failed to control effectively. With the trade restriction in place, this agreement is very likely to work. Climate negotiators would do well to look for other opportunities like this one; opportunities focused on individual gases and sectors for recasting the collective-action problem as requiring coordination. This approach would be complementary to the voluntary-contributions approach taken in the Paris Agreement. However, even these two approaches together are unlikely to be enough to prevent future temperature change.

William Nordhaus (2015) has recently proposed an alternative approach to Paris. This involves enforcing an agreement to limit economy-wide emissions by means of generalized tariffs imposed against nonparties. Under the right circumstances, he shows, this approach could cause all countries to participate. However, his proposal might also prove risky, perhaps sparking a trade war. Further-more, he shows that tariffs lose their power to induce participation as an agreement aims for larger and larger emission reductions. Indeed, to date no one has proposed a politically acceptable approach to limiting (as opposed to simply attenuating) climate change.

## Notes

This article is based on a paper prepared for the Wissenschaftskolleg Fellow Forum Workshop, 'Too Big to Handle: Interdisciplinary Perspectives on the Question of Why Societies Ignore Looming Disasters', Berlin, October 2014. I am grateful to the other participants at this workshop for their comments on my presentation of the paper.

1. For example, four players may contribute €2 each for each of the first five periods and €4 each for each of the last five periods, while the remaining two players contribute €0 every period. In this case, the contributors each get €10 and the noncontributors each get €40.
2. A symmetric mixed strategy equilibrium also exists.
3. An interesting question is whether there is a stronger incentive for countries to avoid 'abrupt and catastrophic climate change' compared to the incentive to avoid 'gradual' climate change when both problems are a prisoners' dilemma. In another experiment, Astrid Dannenberg and I have shown that the prospect of catastrophe does increase contributions, but not by enough to avert catastrophe (Barrett and Dannenberg, 2014b).
4. See www.iy2kcc.org/News20000301.htm [Accessed 5 May 2015].
5. International Y2K Cooperation Center, "Y2K: Starting the Century Right!," Washington, DC: International Y2K Cooperation Center, February 2000, p. 79.
6. For an excellent summary and synthesis of this literature, see Smith (2007).
7. See www.who.int/dg/speeches/2011/malaria_plan_20110112/en/.
8. A 2013 report by the World Health Organization lists nine countries as still allowing the marketing of these drugs: Angola, Cape Verde, Colombia, Equatorial Guinea, Gambia, Sao Tome and Principe, Somalia, Swaziland and Timor Leste. See www.who.int/malaria/monotherapy_NDRAs.pdf.
9. Another 2013 report by the World Health Organization lists 86 manufacturers, of which only 56 have withdrawn these drugs from the market; see www.who.int/malaria/monotherapy_manufacturers.pdf. Most of the 30 companies that continue to market artemisinin-based monotherapies are located in India, Nigeria, China and Pakistan.
10. Using Krasner's (1999) classification, all states, being unable to block imports of resistant pathogens, lack 'interdependence sovereignty'. Many states, being unable to regulate drug distribution, lack 'domestic sovereignty'.
11. See www.armscontrol.org/factsheets/Nuclearweaponswhohaswhat for a nuclear weapons count.
12. In 2010, the US and Russia signed a New START Treaty, which entered into force in 2011. President Obama has declared a global goal of zero nuclear weapons.
13. Nor did Kahn recommend the doomsday machine: 'if one were presenting a military briefing advocating some special weapon system as a deterrent . . . the Doomsday Machine might seem better than any alternative system; nevertheless, it is unacceptable' (Kahn, 1961, pp. 104–105).

## References

Barrett, S. (2003) *Environment and Statecraft: the Strategy of Environmental Treaty-making*. Oxford: Oxford University Press.

Barrett, S. (2006) 'The Problem of Averting Global Catastrophe', *Chicago Journal of International Law*, 6(2), pp. 1–26.

Barrett, S. (2007) *Why Cooperate? The Incentive to Supply Global Public Goods*. Oxford: Oxford University Press.

Barrett, S. (2013) 'Climate Treaties and Approaching Catastrophes', *Journal of Environmental Economics and Management*, 66(2), pp. 235–250.

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

Barrett, S. and Dannenberg, A. (2014a) 'Sensitivity of Collective Action to Uncertainty about Climate Tipping Points', *Nature Climate Change*, 4, pp. 36–39.

Barrett, S. and Danneberg, A. (2014b) '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.

Barrett, S., Lenton, T. M., Millner, A., Tavoni, A., Carpenter, S., Anderies, J. M., et al. (2014) 'Climate Engineering Reconsidered', *Nature Climate Change*, 4, pp. 527–529.

Hardin, G. (1968) 'The Tragedy of the Commons', *Science*, 162, pp. 1243–1248.

Kahn, H. (1961) 'The Arms Race and Some of Its Hazards', in D. G. Brennan (ed.), *Arms Control, Disarmament, and National Security*. New York: George Braziller, pp. 89–121.

Keith, D. W. (2000) 'Geoengineering the Climate: History and Prospect', *Annual Review of Energy and the Environment*, 25, pp. 245–284.

Krasner, S. D. (1999) *Sovereignty: Organized Hypocrisy*. Princeton, NJ: Princeton University Press.

Kriegler, E., Hall, J. W., Held, H., Dawson, R. and Schellnhuber, H. J. (2009) 'Imprecise Probability Assessment of Tipping Points in the Climate System', *Proceedings of the National Academy of Sciences*, 106(13), pp. 5041–5046.

Lenton, T. M. (2011) 'Early Warning of Climate Tipping Points', *Nature Climate Change*, 1, pp. 201–209.

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

Milinksi, M., Sommerfeld, R. D., Krambeck, H.-J., Reed, F. A. and Marotzke, J. (2008) 'The Collective-risk Social Dilemma and the Prevention of Simulated Dangerous Climate Change', *Proceedings of the National Academy of Sciences*, 105(7), pp. 2291–2294.

Nordhaus, W. (2015) 'Climate Clubs: Overcoming Free-riding in International Climate Policy', *American Economic Review*, 105(4), pp. 1339–1370.

North, D. C. (1990) *Institutions, Institutional Change and Economic Performance*. Cambridge: Cambridge University Press.

Ostrom, E. (1990) *Governing the Commons*. Cambridge: Cambridge University Press.

Sagan, C. (1983) 'Nuclear War and Climatic Catastrophe: Some Policy Implications', *Foreign Affairs*, 62(2), pp. 257–292.

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

Sandler, T. (1997) *Global Challenges: an Approach to Environmental, Political, and Economic Problems*. Cambridge: Cambridge University Press.

Sandler, T. (2004) *Global Collective Action*. Cambridge: Cambridge University Press.

Smith, D. L. (2007) 'The Epidemiology of Anitbiotic Resistance: Policy Levers', in R. Laxminarayan and A. Malani (eds), *Extending the Cure: Policy Responses to the Growing Threat of Antibiotic Resistance*. Washington, DC: Resources for the Future, pp. 39–68.

Smith, W. (2000) 'Russia's Nuclear Arsenal: Why the Y2K Bug Didn't Bite', *The Fletcher Forum of World Affairs*, 24(1), pp. 191–197.

Thompson, S. L. and Schneider, S. H. (1986) 'Nuclear Winter Reappraised', *Foreign Affairs*, 64(5), pp. 981–1005.

UNFCCC Secretariat (2015) Synthesis Report on the Aggregate Effect of the Intended Nationally Determined Contributions [online], FCCC/CP/2015/7, 30 October. Available from: http://unfccc.int/resource/docs/2015/cop21/eng/07.pdf [Accessed 22 December 2015].

Velders, G. J. M., Anderson, S. O., Daniel, J. S., Fahey, D. W. and McFarland, M. (2007) 'The Importance of the Montreal Protocol in Protecting Climate', *Proceedings of the National Academy of Sciences*, 104(12), pp. 4814–4819.

## Author Information

**Scott Barrett** is the Lenfest-Earth Institute Professor of Natural Resource Economics at Columbia University.