Luring others into climate action: Coalition formation games with threshold and spillover effects

Valentina Bosetti[*,§], Melanie Heugues[*], Alessandro Tavoni[†]

§ Bocconi University, Milan, Italy and Centro EuroMediterraneo sui Cambiamenti Climatici, Milan, Italy.
* Fondazione Eni Enrico Mattei, Milan, Italy
† Grantham Research Institute, London School of Economics, London WC2A 2AZ, England. Corresponding author: a.tavoni@lse.ac.uk.

Abstract

We explore the prospects of cooperation in a threshold public bad game. The experiment's setup allows us to investigate the issue of effort coordination between signatories and non-signatories to a climate agreement under the threat of a catastrophe. Motivated actors may signal willingness to lead by committing a share of investments to a 'clean' but less remunerative project. The game is parametrized such that the externality cannot be fully internalized by the coalition, so that some effort on the part of the second movers is required if the catastrophic losses are to be avoided. We manipulate both the relative returns of two investments and the extent to which the gains from leadership diffuse to second movers. We find that the likelihood of reaching a sizeable coalition of early investors in the clean technology is higher when the benefits are appropriated by the coalition. Conversely, spillovers can entice second movers' adoption.

JEL classification: C70, C92, Q50

1. Introduction

The global nature of climate change and the fact that long-lived greenhouse gases (GHG) accumulate over time, make it a daunting challenge for humanity. All countries (to varying degrees) are responsible for having contributed to the current atmospheric carbon dioxide concentration; yet, emissions diffuse beyond national borders, and the impacts of the warming climate on a given country are largely decoupled from its responsibility. Inertia in the climate system means climate change will continue even if emissions and atmospheric concentrations are stabilized (Edenhofer et al. 2014). Thus effective mitigation requires almost universal participation and implies costly action to avoid losses. These features can be captured by public bad games: each country benefits from its own emissions, but suffers from damages linked to global emissions. Consequently, incentives to free ride are strong, i.e. each country taken individually has an incentive to continue to emit GHG, letting the others reduce their own polluting activities.

A second dimension of the problem is that climate change, like other socio-ecological processes, may entail sudden transitions from more benign states to harmful ones (tipping points): without sufficient abatement effort, emissions may accumulate beyond a point at which catastrophic and irreversible regime shifts may occur (Alley et al. 2003; Lenton et al. 2008; Kriegler et al. 2009; Lade et al. 2013), putting at stake countries' development and welfare (Stern, 2007).

As a consequence of the above incentives, the economic regulation of the problem of stabilizing GHG emissions calls for the coordination of national environmental policies. At present, the prospects for a meaningful universal agreement achieving such a goal are slim, given the free riding incentives to refrain from reducing emissions, the sovereignty of states and the absence of a supranational authority (Barrett 2006). The ensuing grim prediction is that voluntary action by countries partaking into a self-enforcing international environmental agreement will comprise few parties with ambitious reduction targets (Barrett 1994; Finus, 2001). It appears therefore that the laggards, that is the non-signatories to the treaty, will be instrumental in determining the chances of averting dangerous climate change (Heugues 2014).

In this paper, we investigate experimentally whether the establishment of an institution that commits a coalition of signatories to capping its use of a technology with negative externalities entices others to follow suit, given that the group collectively risks high losses if the amount of

cooperation is insufficient. Based on the premise that in the absence of universal participation an international climate agreement will deliver insufficient commitments to avoid crossing a threshold for dangerous climate change (such as the often mentioned 2^0C warming level), we parametrized the game such that under partial cooperation non-signatories are also needed in order to avoid the catastrophe. That is, except when all players sign up to the treaty, in which case the grand coalition avoids exceeding the threshold, second-movers (while free to choose their investment level) are pivotal to avoiding large losses. This setup is meant to mimic the current state of affairs, with the large gap between the very restricting global objective (abatement between 40% and 70% of global emissions by 2050 compared to 2010 in order to stay below 2^0C warming, according to Edenhofer et al. 2014) and the existing global polluting capacity. This gap means that ambitious mitigation is increasingly being recognized as necessary not only on the part of traditional leaders (mostly developed nations), but also by less committed followers (mostly less developed nations).

While our threshold public bad experiment is framed neutrally, we use the avoidance of dangerous climate change as an illustration of the problem throughout the paper[1]. The existence of a known threshold simplifies the challenge of reaching meaningful agreement in negotiations by transforming the underlying game into one of coordination with two Pareto-ordered equilibria (Barrett and Dannenberg 2012; Dannenberg et al. 2015).[2] Nonetheless, scientific uncertainties and the perceived threat of competitiveness loss from mitigation effort have contributed to widespread unwillingness to take on ambitious action.[3] To help coordination on the safe equilibrium, we consider several mechanisms inspired by the economic literature and test whether they induce reinforcement between different players' strategies. The first mechanism is the introduction of a membership stage in the threshold public bad game through which signatories can signal their commitment: by agreeing to restrict use of the 'polluting' technology,

[1] The subjects were confronted with choices between two investment projects. Compared to a frame that stresses the moral imperative for action (e.g. to reduce global pollution), this choice might induce less overall collaborative behavior (Liberman, Samuels, and Ross 2004). We are interested in treatment effects rather than levels; furthermore, experiments with a neutral frame have the advantage of being less prone to confounding effects that originate from the frame.

[2] Note however that even in the presence of a known threshold with the potential to trigger a catastrophe, coordination can be difficult, especially when parties have different stakes in the game (Tavoni et al. 2011). Furthermore, cooperation has been shown to be significantly lower in public bad games than in the equivalent public good games (Andreoni 1995; Sonnemans et al. 1998)

[3] International negotiations on GHG emissions abatement has been so far unsuccessful in defining national targets that are compatible with the global aspiration to keep global warming within safe boundaries (Rockström et al. 2009).

leading countries may induce non-signatories to follow suit. The second mechanism is cooperation on the 'clean' technology, both among signatories and across the two groups. When technological cooperation benefits only signatories, the game simulates a situation where innovation is appropriable. When technological cooperation also benefits non-signatories, the game mimics a situation in which innovation cannot be fully appropriated. This design allows us to study the effect of coalition formation on the behavior of the fringe, either in the presence or absence of innovation benefits. Lastly, we test the effect of varying the magnitude of the above innovation co-benefits, by introducing treatments with either low or high co-benefits from cooperation. Below we detail the literature supporting these mechanisms.

Some experimental works have shed light on the role of leading by example in facilitating the provision of public goods (Moxnes and Van der Heijden 2003; Levati et al. 2007). Using a public bad experiment, (Moxnes and Van der Heijden 2003) ask whether a leader that takes unilateral action (reducing investment in the public bad) motivates the followers to also reduce investment. They find that leadership has a small but significant effect, provided that the example is sufficiently good (i.e., the leader's investment in the public bad is sufficiently low).[4] The main difference in the game considered here is that we do not appoint the leader exogenously; rather, a coalition of leading players emerges endogenously.

Scaling up climate cooperation through small-scale coalitions has received increasing attention in related theoretical works (Ostrom 2009; Sterner and Damon 2011; Vasconcelos, Santos, and Pacheco 2013; Tavoni 2013; Marchiori et al. 2017). The network diffusion of behaviours and technology adoption may play an important role in fostering cooperation since adoption by one agent often increases the likelihood that others will become aware of the existence of such behaviours and technology and their potential benefits relative to the status quo. Many studies have shown that mutually reinforcing choices lead to accelerated diffusion of a behaviour or the adoption of a technology once a tipping point has been reached (Granovetter 1978; Watts 2002; Weir and Knight 2004). Heal and Kunreuther (2011) focus instead on coordination in games with strategic complementarity, using the concept of a tipping set, which is a subset of agents that can induce all other agents to change from the inefficient equilibrium to

[4] Relatedly, (İriş et al. 2016) find that contributions to a threshold public good drop when the investment decision is delegated to an appointed leader. This effect is attributable to the fact that delegates appear to focus on the lowest contribution level proposed by non-delegates (rather than the highest or average proposals). Hence, negative examples can be detrimental to cooperation.

the efficient one. They argue that international climate agreements have these characteristics and motivate the theory with two frequently mentioned examples of strategic complementarity: the replacement of leaded gasoline with unleaded gasoline and the phasing out of chlorofluorocarbons through the Montreal Protocol on Substances that Deplete the Ozone Layer. Both examples show how unilateral action initiated by a subset of actors (in the United States) prompted others to follow suit immediately afterward.

Innovation and technology cooperation have been frequently suggested as possible ways to end negotiation deadlocks (Carraro and Siniscalco 1995; Barrett 2003; Golombek and Hoel 2004; Barrett 2006) . However, linking coalition efforts with the ancillary benefits stemming from coordinated innovation brings a new externality into the analysis. If cooperation on innovation hinges on partially sharing the associated collective burden (and benefiting from its yield), what about those who were not part of the agreement in the first place? Depending on the nature of the technologies, non-participants could, in principle, be excluded from such benefits—for example, through a system of exclusive property rights. Would this be in the interest of the cooperating group? Given the limited availability of data on this matter, we aim to contribute to this debate through experimental evidence.

The role of spillovers through technology transfers within coalitions and between signatories and non-signatories have been documented by investigating different channels, such as climate policies linkages (see, for example, the work by Dechezleprêtre et al. 2008 and Seres et al. 2009) on technology transfers via the Clean Development Mechanism), trade flows, multinational enterprises, and skilled-labour mobility (Eaton and Kortum 2001; Eaton and Kortum 2006; Keller 2010). Although empirical studies can hardly be definitive on the subject, technological transfers have also been highlighted in the theoretical literature as a mechanism that can in principle generate reinforcement effects between countries' environmental policies (Golombek and Hoel 2004; Di Maria and Van der Werf 2008). These reinforcement effects occur when countries that have not signed an environmental agreement reduce pollution in response to the efforts of an environmental coalition. These studies suggest that in principle, it could be profitable for the coalition to allow non-signatories to benefit from its innovations.

The reviewed body of work suggests that unilateral action by a subset of agents may be able to promote widespread cooperation, the threat of free riding notwithstanding. In the present paper, we test this proposition in the laboratory, by focusing on the role of increasing returns to

coalition size (mimicking increasing returns to scale in the innovation and adoption of clean technologies), as well as on the implications that proprietary and open knowledge policies might have. We employ a threshold public bad game to test how these mechanisms play out in deterring or incentivizing players to join a coalition of early investors and affecting how players that do not join respond to the coalition.

This experiment departs from standard public goods games in at least three ways. First, as in a growing experimental literature that begins with (Bagnoli and McKee 1991), the game employed here includes a threshold, which transforms it into a game of coordination with two equilibria (*disaster avoidance* and *gamble*, as explained in Section 2.1). Second, it includes the possibility of forming a coalition of Stackelberg leaders who invest in a technology that is socially preferable but individually costlier. Third, it includes ancillary benefits to cooperation that may be appropriated by the coalition or diffused to non-members. Furthermore, this study also departs from the existing experimental literature on coalition formation, since it is mainly concerned with the behaviour of the fringe, rather than in the number of signatories. Accordingly, we fix the coalition's investments conditional on the number of signatories (i.e. signatories' only decision is whether to join the treaty, the ensuing investment is pre-determined and known), whereas the remaining players are free to choose their own strategy. This design allows one to study the response of non-signatories to the prior effort of coalition members, in the face of a common objective (avoiding exceeding the tipping point). It is an empirical matter whether the potential benefits from luring the outsiders into action outweigh the potential cost from deterring participation to the coalition.

Our results support the idea that we should move away from agreements seeking universal participation on emission reduction targets in isolation from other policies, since it exposes cooperators to the 'tyranny of free riders,' who refuse to undertake sufficiently ambitious efforts for the group to avoid going over the tipping point. This well-known negative result is alleviated when: (i) there exist sufficiently large returns to signing up to a coalition and the co-benefits are appropriated by it; or (ii) the fringe can partake in the benefits generated by the coalition, which acts as stimulus for (partial) cooperation by the second-movers. This finding casts new light on the problem by highlighting the game-changing potential of linking an environmental agreement with technological cooperation, as well as the strategic implications of opening or restricting access to the new technology.

Before detailing the experimental design in Section 3, we describe the main features of the game in the next section. Section 4 discusses the main findings of our experiments and Section 5 draws some conclusive remarks.

2. The game

We start by presenting the game in the absence of coalition formation; this will be useful as a baseline. We then provide the stages of the game to shed light on how coalition formation and technological cooperation can help coordination. Lastly, we solve the game by backward induction.

2.1 The threshold public bad game

Consider N symmetric players taking part in a threshold public bad game. Each player has an initial endowment e and decides how much to allocate between the high-return but socially costly Project A (the public bad, henceforth, 'A') and the lower-return Project B, which does not cause negative externalities (henceforth, 'B'). The endowment is thus split between x_A and $x_B = e - x_A$. Investing in A (B) gives a private return of r_A (r_B). Returns on A are thus larger than on B, $r_A > r_B > 0$, but A has also a negative external effect: each unit invested in A yields a negative return of c_A to all players.

In addition to this traditional negative externality, the group's aggregate investment determines whether a tipping point has been reached. The threshold *T* represents the maximum safe collective investment in A that is compatible with full enjoyment of the private earnings. To make the problem relevant, this safe level has to lie below the maximal public bad investment capacity (Ne). Players thus retain their earnings with certainty (*disaster avoidance*) if $Nx_A \leq T < Ne$; otherwise, they are left with $q \in [0, 1)$ of their private earnings (*gamble*) with probability p.

A player's expected payoff then takes the form:

$$\begin{cases} \pi(x_A, x_B) = r_A x_A + r_B x_B - c_A \sum x_A, & if\ X_A \leq T \\ \pi(x_A, x_B) = (1-p)[r_A x_A + r_B x_B - c_A \sum x_A] + pq[r_A x_A + r_B x_B - c_A \sum x_A], & if\ X_A > T \end{cases} \quad (1)$$

where N, $e \in R^+$, x_A and $x_B \in [0, e]$, and $c_A < r_A - r_B < c_A N$. The first inequality means that A's private net return is larger than B's, i.e. $r_A - c_A > r_B$; the second means that the individual

opportunity cost of investing in B, is lower than the social marginal cost of pollution, $c_A N$. The latter inequality is consistent with the existing empirical evidence (Stern, 2007; IPCC, 2014[5]).

The social optimum entails that all players refrain completely from investing in A. In this case, each subject gets $\pi(0, e) = r_B e$. However, this is not an equilibrium, as each player has an incentive to deviate. By increasing x_A by one unit, any individual can get $\pi(1, e-1) = r_A + r_B(e-1) - c_A$ (while the others get $\pi(0, e) = r_B e - c_A$). As long as A's net return is larger than B's, the deviation pays off. Hence, a dilemma arises since each individual strictly prefers to invest everything in A, assuming that all others refrain from investing. The more subjects follow this line of reasoning, the lower everyone's expected payoff is (because of the gradual negative externality term $c_A \sum x_A$ as well as the increased likelihood of crossing the threshold). Risk-neutral players will either coordinate to avoid the tipping point (by collectively investing exactly T in A), or disregard the externality and invest all their endowment in A. These two symmetric Nash equilibria correspond to $\overline{x}_A = T/N$ and $\underline{x}_A = e$ and are Pareto-ordered. We denote by $\overline{\pi}$ the payoff associated with disaster avoidance ($\overline{x}_A = T/N$) and call $\underline{\pi}$ the payoff associated with the dominated Nash equilibrium and obtained investing the entire endowment in A ($\underline{x}_A = e$). It can be easily checked that $\overline{\pi}(\overline{x}_A, e - \overline{x}_A) > \underline{\pi}(e, 0)$ is always true.

2.2 Coalition formation with technological cooperation

To capture the element of leadership, we introduce a membership stage where players can form a coalition of committed leaders. Here, signing an agreement means adhering to a pre-specified investment strategy that is conditional on the number of signatories. In the ensuing stage, non-signatories choose freely how much to invest in A.

It is worth stressing that we depart from the majority of experiments involving coalition formation with respect to two dimensions (Kosfeld et al. 2009; Burger and Kolstad 2009; McEvoy et al. 2011; Cherry and McEvoy 2013; Dannenberg et al. 2014; McEvoy et al. 2015). First, we require signatories to agree to curtail investments in A only partially and according to a pre-defined strategy. Second, we don't restrict non-members' behaviour as usually done, since the behaviour of the fringe in response to partial coalitional cooperation is the main focus of our approach.

[5] For further details, see (Kolstad et al. 2014, 851)

In the extreme case where the grand coalition forms, exactly T is invested in A, thus guaranteeing the achievement of the preferable equilibrium. For smaller coalitions, achieving the disaster avoidance equilibrium $X_A = T$ requires some restraint in the investment in A by the fringe as well. This setup thus allows us to investigate the issue of coordination between members and non-members to a climate agreement under the threat of an impending catastrophe. The rationale is that more motivated countries may be willing to lead in the costly transition to net-zero emissions, but some effort will still be required on the part of second movers.[6]

Formally, we identify by s the number of members joining the coalition. For any $s \in [2, N-1]$, each coalition member is required to curtail investment in A to an amount that is strictly less than T/N; $x_A^s = X_A^s/s < T/N$.[7] That is, the game is designed such that, except when $s = N$ (in which case each signatory to the grand coalition invests exactly T/N), members will restrict investment in A beyond what a symmetric burden sharing would entail. This design has two effects: (1) it facilitates threshold coordination by giving the second-movers some room to manoeuvre compatibly with not exceeding T; (2) it triggers an incentive to free-ride for non-members.

While one may question requiring this altruistic choice on the part of signatories, the goal is to test whether this kind of leadership (exerted by the coalition through generous restriction in public bad investment) induces a positive response by the fringe even in the presence of the ensuing free-riding incentives. There is ample anecdotal evidence of this kind of 'going the extra mile' behaviour by committed parties to environmental treaties, an example being the above-average commitments by the European Union both in the Kyoto Protocol and Paris Agreement. Arguably, this behaviour is aimed at enticing some degree of cooperation by laggard countries that are less willing or capable of contributing (Clark, 2016). The case of the EU's stance on climate negotiations is exemplary of the benefits of testing conflicting incentives in a controlled laboratory setting which avoids the counterfactual problem and can provide us with an indication of which effect is likely to dominate in determining the behaviour of the fringe.

[6] Inspection of the Nationally Determined Contributions (NDCs) submitted as part of the Paris Agreement suggests that the commitments, even if they were fully implemented, are incompatible with achieving the target of keeping the global temperature increase below 2°C. Such an objective will only be feasible if more reluctant countries (for instance high emitters such as India) follow suit with substantial divestment from fossil fuels.

[7] Note that under this restriction, average investment by non-members above T/N can still be compatible with avoiding the probabilistic loss triggered by exceeding T; see Table A2 in Appendix 2 for further details.

Following the membership stage, the $N - s$ non-signatories who opted not to join the coalition (henceforth indexed n), are free to choose their investment in light of the information on the size and aggregate investment of the coalition. The resulting total investment in A can then be expressed as $X_A = X_A^s + X_A^n$. This determines the group's standing with respect to the threshold *T* and the externality cost $c_A X_A$, both of which affect each individual's payoff.

Given our interest in studying the role of co-benefits in promoting or deterring cooperation, we also consider the case when the returns to B increase with the size of the coalition. This modification of (1) effectively reduces the profitability gap between the two investments, as detailed in (2). Inspired by the literature on multi-issue bargaining (Schelling, 1960), we assume that, by curtailing investments in A, members of the coalition also increase the productivity of B. The adoption of new technologies typically entails economies of scale and co-benefits, and the rationale of our setup is that by joining forces, signatories leverage the coalition size to increase productivity, thus reducing the return wedge between the two technologies. The larger s, the greater the co-benefit, i.e. the larger the positive externality on B. $r_B(1 + sI)$ is the increased return to B resulting from a coalition of size *s*, where $I \in [0,1)$ is the percentage rate of technological improvement.[8]

In order to account for this innovation externality, the payoff to its beneficiaries takes the following modified form:

$$\begin{cases} \hat{\pi}(x_A, x_B) = r_A x_A + r_B(1 + \boldsymbol{sI})x_B - c_A X_A \; if \; X_A \leq T \\ \hat{\pi}(x_A, x_B) = (1 - p + pq)[\, r_A x_A + r_B(1 + \boldsymbol{sI})x_B - c_A X_A] \; if \; X_A > T \end{cases} \quad (2)$$

We investigate two alternative setups to study the implications for the fringe of having technological cooperation among signatories. In the first, the *no spillover* case, the positive externality is appropriated by coalition members only, with non-signatories' payoffs given by (1). In the second setup, the *spillover* case, we assume that this positive externality diffuses to the fringe, whose payoff then follows equation (2).

We now discuss the parameterization of the treatments that were run in the laboratory. The equilibrium solution of the two-stage game is given online in Appendix 1.

[8] To keep the social dilemma, we impose $r_B(1 + NI) < r_A - c_A$; that is, even when all subjects cooperate, the net return to A remains larger than the increased return to B.

3. The experimental design

3.1 Parametrization and treatments implemented

Groups of $N = 7$ subjects face a game described by means of a neutral language. Each player is endowed with $e = 50$ experimental currency units (1 ECU = 0.05 Euro), which are to be entirely allocated between the two investment projects A and B. Each unit invested in A yields an individual return $r_A = 10$ and causes a cost $c_A = 1$ to each group member; B yields a lower return $r_B = 6$ but carries no external cost, $c_B = 0$. The threshold is set at $T = 105$ ECU $=$ $30\%\ N * e$, meaning that for a group to avoid probabilistic losses, it must limit its collective investment in A to at most 30% of the total endowment (or, equivalently, invest at least 70% in externality-free B). Otherwise, all subjects in a group face a 50% probability of losing their earnings, $p = 0.5$ and $q = 0$.

With regard to the increased competitiveness of B resulting from coalitional investments in it, we test the following cases: $I = \{0\%;\ 2\%;\ 7\%\}$. We refer to *I* = 2% as the condition with *low innovation returns* and to *I* = 7% as the case with *high innovation returns*.[9] We also manipulate whether the returns to innovation are appropriated by coalition members only or benefit the fringe as well (the *spillover condition*).

We test for the effects of four conditions, yielding the five treatments (and the control) given in Table 1. The threshold public bad game described in Section 2.1 serves as the benchmark (T0). It consists solely of the investment decision stage, where players simultaneously and independently choose their investment in A (the remainder, if any, is invested in B). T1 captures the implications of the addition of a membership stage with coalition formation, while the remaining four treatments differ in terms of the returns to innovation (low returns in T2 and T3 and high returns in T4 and T5) and who the beneficiaries are (in T2 and T4 only coalition members benefit from the increase in r_B, whereas in T3 and T5 every player can benefit). The treatment features are summarized in Table 1.

Table 1. Features of the different treatments.

	COALITION	***I* = 2%**	***I* = 7%**	**SPILLOVER**
T0 (*Threshold Public Bad Game*)				
T1 (*Coalition Only*)	✓			
T2 (*Coalition & Low Innovation*)	✓	✓		

[9] Recalling (2), for positive *I* and coalition size *s*, the return to B increases to $r_B(1 + sI)$.

T3 (*Coalition & Low Innovation with Spillover*)	✓	✓		✓
T4 (*Coalition & High Innovation*)	✓		✓	
T5 (*Coalition & High Innovation with Spillover*)	✓		✓	✓

To recap, coalition treatments (i.e. T1 to T5) are such that in the first stage subjects decide whether or not they want to become a member of the coalition. They do so knowing that signatories are bound to pre-determined investments in A that are known to all players before the membership decision and are conditional on the size of the coalition that is formed (as detailed in Table 2). As a consequence, in stage 2, members make no decision. Non-members can instead freely choose their investment in A, after having found out what the size of the coalition is, as well as the implied investment in A by the members, the corresponding total amount that can be invested in A before reaching the threshold, and the ensuing (maximal) symmetric individual investment in A compatible with not exceeding the threshold.

Once information about investments by non-signatories is collected, each subject is informed of the resulting aggregate investment in A, whether the threshold has been crossed, and her/his final payoff (conditional on the 50% probability of loss for instances where the group exceeded the threshold).[10]

Table 2 below reports the information given to subjects at the membership stage for treatment T1 in order to inform their membership decision. This includes the exogenously fixed investment in A by each signatory (x_A^S) and how it varies with the coalition size: the smaller the coalition, the more effort is required from each coalition member in terms of constraining investment in A (see Appendix 2 for further details). The table also reports, for each coalition size, the remaining amount that non-members can invest in A without exceeding the threshold, both at the group (penultimate row) and at the individual level (assuming symmetric behaviour, see bottom row).[11]

Table 2. Information provided to players during the game in T1 (Coalition Only)

	Number of signatories (s)					
	7 (all)	**6**	**5**	**4**	**3**	**2**

[10] When the threshold was crossed, a virtual coin was tossed to determine the outcome, and the effective payoff was communicated to the group.

[11] In treatments T1 to T5, subjects were informed that i) a coalition only forms if at least two participants in the group signed up; ii) the members of the coalition cannot guarantee that the sum of all investments in A will not exceed 105 ECU unless all seven join the coalition.

Investment in A by each member (x_A^s)	15	13	11	9	7	5
Aggregate investment in A by the coalition (X_A^s)	105	78	55	36	21	10
Amount left to be invested before reaching 105 ECU (T-X_A^s)	0	27	50	69	84	95
Corresponding symmetric individual investment not to exceed 105 ECU for non-members ($\frac{T-X_A^s}{N-s}$)	0	27	25	23	21	19

For treatments where the implications of innovation were tested and the return wedge between A and B was reduced proportionally to coalition size (T2-T5), the information on the increased return to B was provided at the membership stage, in addition to the information presented in Table 2. To save space, in Table 3, we lump together different information that was provided in different treatments. Depending on the treatment, participants were informed that the returns would apply to coalition members only or to all players.

Table 3 – Returns to Project A and B under different treatments and coalition sizes

	Number of coalition members						***No Coalition***
	7 (all)	**6**	**5**	**4**	**3**	**2**	**0**
Gross return from Project A (T0 –T5)	10	10	10	10	10	10	10
Return from Project B to all (T1)	6	6	6	6	6	6	6
Return from Project B to members only (T2) [to everybody in T3]	6.84	6.72	6.6	6.48	6.36	6.24	6
Return from Project B to members only (T4) [to everybody in T5]	8.94	8.52	8.1	7.68	7.26	6.84	6

3.2 Hypotheses

In this section, we derive four hypotheses, assuming symmetry and risk neutrality. The hypotheses are based on the following theoretical predictions for the different treatments (the computations are detailed in Appendix 2):

- In T1, T2, T3 and T5, the stable coalition will comprise two signatories: $s^* = 2$. Non-signatories will either coordinate on the disaster avoidance equilibrium or invest maximally in A (yielding the gamble equilibrium). The ensuing investments in A for the signatories and the non-signatories are respectively $(X_A^s = 10; T - X_A^s = 95)$ and $(X_A^s = 10; (n-s)e = 250)$.

- In T4, $s^* = N$, yielding the disaster avoidance equilibrium with $X_A^s = 105$.

Thus, with the exception of T4, the theory does not identify which equilibrium will be selected by the non-signatories, as expected in a coordination game. Laboratory evidence will help us to determine which equilibrium is more likely in a given treatment. However, based on the above

predictions, on the incentives to join the coalition (Table A1), and on the incentives to coordinate on disaster avoidance for non-members (Table A2), we can state the following hypotheses:

> *Hypothesis 1 (Coalition size)*: A coalition will form in all the treatments featuring a membership stage (T1-T5). Among those, *s* will be the largest in T4, and very small in the other treatments.

The first hypothesis simply restates the above reasoning. As a side note, the incentives to join a coalition are lowest in T1, relative to the incentives in treatments featuring co-benefits of early investments in B (T2–T5), as detailed in Table A1.

We now turn to the behaviour of the non-signatories and the issue of equilibrium selection. Based on Hypothesis 1 we restrict attention to T1, T2, T3 and T5, since T4 is expected to induce universal agreement in the first stage. While one cannot formulate a precise prediction about which of the two equilibria will be selected by the fringe, we expect that the collective incentive to coordinate on disaster avoidance will induce at least some of the non-signatories to curtail their investment in A.

> *Hypothesis 2 (Fringe behaviour)*: In treatments T1, T2, T3 and T5, non-signatories are more likely to coordinate on disaster avoidance than to select the gamble equilibrium, since avoiding to exceed *T* (just) yields higher expected payoff than investing maximally in A and taking a gamble.

Non-signatories are confronted with two symmetric equilibria. One ensures that the group (just) avoids the tipping point, with each non-signatory investing $\frac{T-X_A^S}{N-s}$ in A; in the other, the fringe takes a gamble and each non-signatory invests all of e in A. Hypothesis 2 rests on the assumption that risk-neutral non-signatories will play according to the subgame perfect Nash equilibrium that guarantees the highest expected payoff given *s*. In fact, as shown in Table A2, a player is always better off restricting investment in A so as to avoid the (probabilistic) losses from crossing the threshold. For a given *s*, the incentives for non-signatories are identical across treatments without spillovers (T1, T2, and T4). In T3 and, to a greater extent, in T5, non-signatories have a larger incentive to coordinate on the disaster avoidance equilibrium as the relative return to B increases. Furthermore, in spillover treatments (especially in T5), the

incentives for non-signatories to limit investments in A are not only higher than in other treatments but are also less sensitive to coalition size—that is, they decrease less in *s*.

On the other hand, while investing maximally in A is risky due to the ensuing probabilistic loss, coordinating on loss avoidance also entails risks, in the sense that one risks being the 'sucker' who curtails investment in A while coordination fails due to one or more non-signatories choosing to invest e in A. Therefore, we rely on empirics to validate (or refute) our hypothesis on equilibrium selection.

Note that the magnitude of the average investment level by non-signatories, which we will call x_A^n, may at first glance be misleading. As pointed out when discussing Table 2, a larger *s* will allow larger investments in A by each remaining non-member that are compatible with *T*: the larger the coalition, the more a non-member can individually invest in A without crossing the threshold. Hence, higher levels of x_A^n do not necessarily mean that non-signatories are failing to coordinate on the safe equilibrium. To avoid confusion, in the following, we will evaluate the degree of cooperation of the fringe by referring to the distance *d* between the average investment level by non-signatories and what they should invest symmetrically in order to coordinate on the threshold, $d = x_A^n - \bar{x}_A^n$, where $\bar{x}_A^n = \frac{T - X_A^s}{N - s}$. The smaller this distance (when non-negative), the closer the fringe is to collectively limit investment in A to the safe level $T - X_A^s$.[12]

The next hypothesis builds on the reasoning developed in Hypothesis 2 about the behaviour of the fringe, but it focuses on the incentives from the positive innovation externality introduced in the spillover treatments.

> *Hypothesis 3 (Spillover effect)*: Spillover of the innovation co-benefits outside the coalition increases the incentive to curtail investment in A for non-signatories. Accordingly, the fringe will be most cooperative in T5, due to higher incentives to coordinate on disaster avoidance, followed by T3. T1 and T2 will feature the least cooperation by the fringe.

Hypothesis 3 also excludes T4, since it is expected to induce a universal agreement in the first stage. From Hypothesis 1 we know that a small coalition is expected to form in all remaining

[12] $d = 0$ captures a situation of perfect coordination on disaster avoidance by non-signatories, while $d < 0$ would be recorded if non-signatories "overshot", in the sense that they went too far in curtailing investment in A, compared to the optimum. In the experiment, we expect subjects to respond to incentives with some degree of fuzziness, compared to the sharp predictions entailed by equilibrium behaviour, as customary in the lab. Hence, d should be a useful indicator of non-signatory cooperation.

coalition treatments ($s^* = 2$). Thus, any difference in the outcome of the game in treatments T1, T2, T3 and T5 hinges on the behaviour of non-signatories. We have a clear prediction in terms of investment in the polluting technology for most of these treatments: d will be closest to 0, the value guaranteeing disaster avoidance, in T5 (with I = 7% and spillovers), followed by T3 (with I = 2% and spillovers), followed by T1 and T2 (where spillovers are absent).

The joint implications of the above hypotheses for the likelihood of averting disaster are summarized below.

> *Hypothesis 4 (Likelihood of disaster)*: The threshold T will be crossed by most groups in T0, followed by both T1 and T2, then T3, followed by T5. Fewest groups will fail to avert disaster in T4.

By comparing T0 with the other treatments, we are able to assess whether the opportunity to form a coalition changes the aggregate investment behaviour. Based on the literature (e.g. Andreoni 1995; Sonnemans et al. 1998; Moxnes and Van der Heijden 2003; Levati et al. 2007; İriş et al. 2016), we expect that the baseline threshold public bad treatment T0 will feature highest total investment in A, because there is no coordination institution in place. Given Hypothesis 1, we expect that in T4 the size of the coalition will be largest, yielding a higher likelihood of disaster avoidance. The other treatments are in between, according to hypotheses 2 and 3. Specifically, Hypothesis 4 allows us to rank the likelihood of disaster as higher in both T1 and T2 than in T3, followed by the high innovation and spillover treatment T5.

In order to test the above hypotheses, we resort to empirical evidence from the laboratory.

3.3 Laboratory Setup

The experiment was conducted between May 2013 and February 2014 at the BELSS laboratory at Bocconi University (Italy). We recruited 434 undergraduate and graduate students from Bocconi University (respectively 70, 84, 63, 70, 77, and 70 subjects for Treatments 0, 1, 2, 3, 4, and 5), corresponding to 9–12 independent group observations per treatment. No subject participated in more than one session. Sessions lasted between 60 and 90 minutes and were run on linked computer terminals; students sat in cubicles and could not see the other participants.

All sessions were run using z-Tree software (Fischbacher 2007). Subjects earned 13.20 Euros on average (including a five Euro show-up fee).[13]

At the beginning of a session, instructions with a neutral frame were provided to the subjects (the context and language of the experiment were abstracted from environmental or other interpretations). Before starting the experiment, subjects completed a comprehension questionnaire to ensure that they fully understood the procedures and the payoff implications of their choices.

Each session consisted of two practice rounds and eight independent rounds; subjects thus played the game described above ten times. They were informed that only the eight independent rounds would be considered in determining the final pay-out. At the beginning of each round, subjects were randomly assigned to groups of seven and were given an endowment $e = 50$ ECU. They were not aware of whom they were grouped with, and no subject was matched up with the same six co-players for more than a single round.[14] At the end of each round, participants were informed of their potential earnings for the round, given their choices and the choices made by the other six players in the group.

At the end of the experiment, subjects were paid according to one randomly selected round (out of eight). Payments were made in cash at the end of the experiment. Since subjects were informed that the round to be selected for payment was randomly determined and could be any of the non-practice rounds, it is reasonable to expect that they played as if each round was payoff-consequential. See Appendix 5 for the instructions given to participants to T5.

4. Results

In line with the hypotheses derived in Section 3.2, we explore the experimental results through three indicators of group performance: the size of the coalition (s), the frequency of threshold-crossing (T^+), and the total investment in the public bad (X_A), which determines the group's distance from the threshold (the extent to which investments in A are below or above *T*). The last metric is relevant because it captures the gradual component of the external costs of investing in A and allows us to explore the role of fringe behaviour. In Table 4, we report the

[13] We used cloud-based participant management software by Sona Systems. The subjects' average age was 22; at the time of the experiment, students were pursuing economics, finance or management degrees. 75% of the subjects were undergraduate students, while the remaining were masters or PhD students; 77% of the subjects were Italian.

[14] This is not a perfect stranger design since a subject in a session could be matched to another for more than one round. Nevertheless, subjects knew that they would not be re-matched in the same group.

summary statistics for these key indicators across all treatments. The first four rows of the table recall the basic assumptions for each of the treatments, while the bottom six report the statistics concerning the performance indicators. All numbers reported are averages over all periods.

Table 4. Descriptive statistics of the results

ASSUMPTIONS	**T0 (Threshold Public Bad Game)**	**T1 (Coalition Only)**	**T2 (Coalition & Low Innovation)**	**T3 (Coalition & Low Innovation with Spillover)**	**T4 (Coalition & High Innovation)**	**T5 (Coalition & High Innovation with Spillover)**
Number of stages	1	2	2	2	2	2
Coalition effect on costs	N/A	None	Internal only	Internal and External	Internal only	Internal and External
Repetition of game with random reassignment	8	8	8	8	8	8
***I*: Increased return to B per member ($s \geq 2$)**	N/A	0%	2%	2%	7%	7%
DESCRIPTIVE STATISTICS	**T0**	**T1**	**T2**	**T3**	**T4**	**T5**
s ; $N - s$ (coalition size; fringe size)	N/A	1.7 ; 5.3	2.5 ; 4.5	1.8 ; 5.2	4.3 ; 2.7	2.5 ; 4.5
x_A^s (Average signatory investment in A)	N/A	7.1	7.4	6.6	10.4	8.3
x_A^n (Average non-signatory investment in A)	27.3	29.2	30.1	24.0	26.6	26.4
$\bar{x}_A^n$ (Symmetric non-signatory investment in A yielding $X_A = T$)	15	17.5	19.2	17.9	22.3	18.7
d ($= x_A^n - \bar{x}_A^n$)	12.3	11.7	10.9	6.1	4.3	7.7
X_A (Total investment in A)	191	166	154	137	117	139
T^+ (Groups that exceeded T)	89%	83%	81%	68%	56%	59%
How far above T	82%	58%	47%	30%	11%	32%
Frequency of coalition formation (% of groups with $s \geq 2$)	N/A	57%	84%	64%	99%	79%

These descriptive statistics provide a preview of treatment effects in this game. In line with our first hypothesis, the number of signatories is largest in T4, where the average coalition size is 4.3. In addition, *s* is close to the predicted value of 2 in all other coalition treatments.

According to Hypothesis 3, *d* should be smallest in T5, followed by T3, and largest in T1 and T2. Contrary to this, the experimental treatment that elicited most cooperation, as captured by *d*, is T3, with an average value of 6.1, followed by T5 with $d = 7.7$. Consistent with expectations, we find a much higher value of *d* in both T1 and T2 (respectively 11.7 and 10.9). Thus, while the exact ranking is not as expected according to theory, the qualitative essence of the spillover effect holds in the lab: treatments featuring spillovers catalyse greater levels of cooperation by the fringe than treatments where innovation benefits are appropriated by the coalition.

The last testable prediction comes from Hypothesis 4, which stipulates that T^+ should be highest in T0 (89%), followed by both T1 and T2 (83% and 81%, respectively), T3 (68%), T5 (59%) and finally T4 (56%). The ranking based on the experimental data is in complete agreement with the expected one, although not all differences are significant (see Figure 1 below).

The remainder of this section is devoted to the analysis of the mechanisms underlying these results.

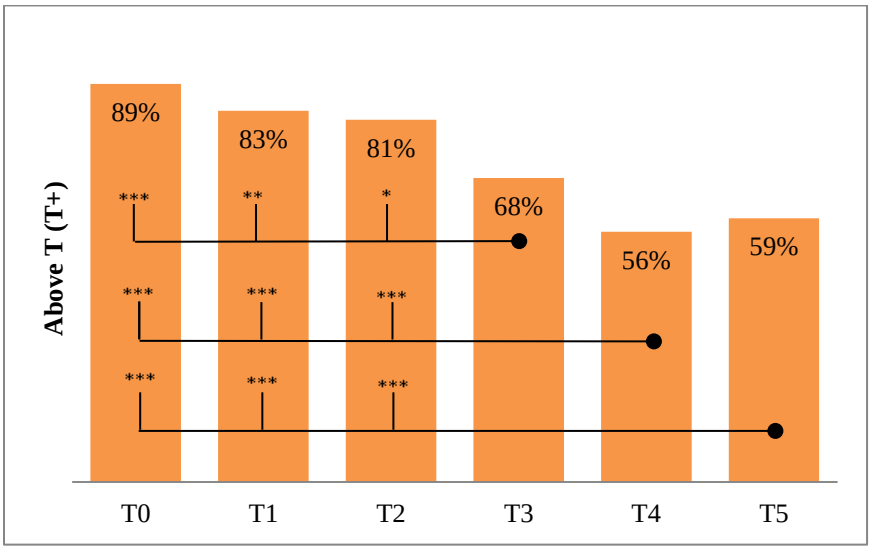

Figure 1: Percentage of groups that exceeded the threshold *T*. The lines indicate statistical differences across treatments (*: $p < 0.01$, **: $p < 0.05$, *: $p < 0.1$). Further analysis is given in Appendix 3.**

To summarize the treatment effects, we map differences using several metrics in Table A3 (Appendix 3). In addition to total investment in A (X_A), failure to avoid crossing the threshold (T^+) and coalition size (s), we also compare total and average fringe investment in A (X_A^n, x_A^n, respectively), in order to get a sense of the relative implications that different incentives have on non-signatories' behavior. Below are the main conclusions that can be derived from this table.

Result 1 (T0):

Given a collective goal, subjects are not able to coordinate on the safe equilibrium in the absence of an institution to signal leadership: in most cases (89%), total investment in Project A is well above the threshold (105 ECU).

The level of cooperation observed here is rather low, as the rate of provision failure observed here is at the low end of the spectrum among those recorded in related threshold public goods experiments (e.g. Milinski et al. 2008; Tavoni et al. 2011; Barrett and Dannenberg 2012; Dannenberg et al. 2015)[15]. Three reasons may explain the discrepancy: the very demanding threshold we set here (-70% of the total available investment in A); the number of players in a group ($N = 7$ is relatively large and makes coordination more challenging than in smaller groups); lastly, we used of a public bad setting, which has been found to induce less cooperation, compared to a public good frame (Andreoni 1995; Sonnemans et al. 1998). Although these differences are interesting per se, this treatment is not the main focus of our study. Rather, we introduced it as an internally consistent benchmark for comparison with the other treatments.

Result 2 (T1):

Absent other mechanisms, leadership by a (small) subgroup of players is not able to entice cooperation from non-signatories. The option to form a coalition is, on its own, ineffective in inducing the fringe to curtail investments in A sufficiently to reduce the likelihood of disaster significantly.

Comparing T0 with T1, we find that the option of signalling leadership triggers a significant reduction in total investment in A, but this is not sufficient to significantly reduce the frequency of threshold crossing (which happens 83% of times in T1). In a discrete public bad setting, this finding is reminiscent of Carraro and Siniscalco (1993) and Barrett (1994)'s dismal conclusion that the option to form a coalition with voluntary participation leads to modest improvements only. This pessimistic result has been confirmed experimentally for linear public goods settings.[16] As we will see below, coalition formation becomes much more consequential when it interacts with the other conditions.

[15] (Milinski et al. 2008 and Tavoni et al. 2011) differ from the present study in that they employ a dynamic setting where multiple contribution trajectories are compatible with providing the public good of disaster avoidance, since the threshold applies to aggregate contributions over 10 rounds. In their closest treatment to T0, with symmetric wealth and no communication,(Milinski et al. 2008) report a 90% provision failure rate, while (Tavoni et al. 2011) find 50% failure. In contrast, (Barrett and Dannenberg 2012) report a 20% failure to avert catastrophe in a one-shot threshold public good game, although cooperation drops sharply in conditions with uncertainty on the location of the threshold, as in (Dannenberg et al. 2015).

[16] See (Dannenberg et al. 2014) for a recent experiment and the references therein for earlier experimental work on coalition formation.

Result 3 (T2 and T4):

Adding an innovation co-benefit that accrues to coalition members increases the likelihood of disaster avoidance.

As we saw, and in line with Hypothesis 1, if the innovation benefit is sufficiently high (I = 7%) and it can be fully appropriated by the signatories as in T4, the incentives to sign up to the treaty are high. Contrary to the theoretical prediction, however, participation is partial even in T4, which averaged s = 4.3 instead of 7. In T2 average s = 2.5, slightly above the predicted value of 2. How did the fringe respond to the larger leading group in T4 relative to T2? Is leadership by example effective in inducing cooperation by the followers?

In T4 investment in A is lower than in T2. More interestingly, the average fringe investment is significantly lower in T4. This suggests that for leadership to be effective, a critical mass is necessary. The resulting implication is that, overall, threshold crossing is significantly lower in T4 compared to any of the other treatments investigated so far (T0, T1 and T2).

Result 4 (T3 and T5):

Although innovation spillovers outside the coalition decrease coalition size (comparing T3 with T2 and T5 with T4), the total investment in A is nonetheless reduced.

Indeed, spillovers reduce the average fringe investment in A, keeping in check X_A in spillover treatments. The above set of results is summarized by the linear regressions reported in Table 5.[17]

Table 5. OLS regression on group-level data.

	S	T^+	T^+	X_A^n	X_A^n	x_A^n	x_A^n
T2	0.587***	-0.0278	0.0675	-18.88**	1.072	0.500	0.915
	(0.195)	(0.0690)	(0.0620)	(8.861)	(5.963)	(1.136)	(1.138)
T3	0.0146	-0.158**	-0.156***	-29.79***	-29.29***	-5.182***	-5.171***
	(0.189)	(0.0670)	(0.0596)	(8.604)	(5.727)	(1.098)	(1.090)
T4	2.285***	-0.277***	0.0945	-81.91***	-4.221	-1.745	-0.0350
	(0.184)	(0.0653)	(0.0682)	(8.388)	(6.559)	(1.081)	(1.253)
T5	0.677***	-0.246***	-0.136**	-35.83***	-12.81**	-3.228***	-2.698**

[17] We cluster errors by individuals to account for repeated observations. Note that we utilized a (quasi) stranger, between-subject design in order to limit learning effects.

	(0.189)	(0.0670)	(0.0605)	(8.604)	(5.817)	(1.098)	(1.108)
s			-0.162***		-34.00***		-0.782***
			(0.0157)		(1.507)		(0.296)
Constant	2.494***	0.622***	1.027***	125.2***	209.9***	25.34***	
	(0.206)	(0.0730)	(0.0758)	(9.381)	(7.287)	(1.211)	
Round Dummy	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	416	416	416	416	412	412	412
R-squared	0.349	0.108	0.296	0.262	0.6740	0.150	0.165

Note: The reference treatment is T1. T0 is excluded from the analysis.

Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

The first column shows the effects on coalition size, whereas the second and third columns predict threshold-crossing and the last two columns report the results for the total and average fringe investments. The coalition size is significantly larger in treatments T2, T4, and T5 than in T1 (at the 1% significance level), but it is not larger in T3 (T0 is excluded because it does not feature coalition formation). However, the frequency of threshold crossing is significantly lower in T3 thanks to the behaviour of the fringe. Considering the total and average investments by non-members, it is clear that self-restraining behaviour by the average fringe member is highest in T3 (the last column in Table 5). T4, on the other hand, implies the largest coalition size, and hence the minimum aggregate investment in A by the fringe. As noted above, this follows from increased participation in the coalition rather than from cooperative fringe behaviour (whose average investment in A remains largely unchanged).

In Appendix 4 we further search for individual characteristics that might influence the individual amounts invested in A and the choice to join the coalition (using a random-effect regression).

5. Discussion

We have empirically explored the prospects of cooperation in a threshold public bad game. This setup allows us to investigate effort coordination between signatories to an international environmental agreement and non-signatories who refrain from making early commitments to curtail investments in a polluting technology, but are also instrumental to the avoidance of an impending catastrophe. As is the case in the real world, more motivated actors may be willing to lead in the simulated negotiations, but it is unlikely that they will be able to solve the externality problem and avoid dangerous climate change without some effort on the part of second movers.

Our analysis suggests two possible situations. In the first scenario, the potential benefits of innovation generated within agreements to foster early investments in a clean technology are modest. In this case, our experiment suggests that negotiations are likely to result in a coalition that is not sufficiently large to be pivotal in inducing the fringe to curtail investment in the polluting technology. Leveraging the effort of second movers by fostering clean technology uptake by the fringe is recommended in such a scenario.

In the second scenario, the prospects of the new technology are sufficiently good to attract a pivotal number of participants into action. The fringe also reacts proactively to their own diminished burden even when the co-benefits do not diffuse to them, perhaps because the target is within sight. To promote this process, it makes sense to commit e*x ante* to some form of appropriation of the knowledge created within the coalition.

Specifically, in order to disentangle the push and pull factors behind the incentives to sign environmental agreements (and, more broadly, behind technology adoption), we have introduced several modifications to the threshold public bad game employed in the baseline treatment. These modifications capture some realistic features of current negotiation platforms and may ease the problem of equilibrium selection. Namely, we have incrementally added the following: (i) a membership stage at which motivated investors in the cleaner technology can lead by example and (partially) correct the externality; (ii) first-mover advantages of differing magnitudes, which increases the competitiveness of the cleaner technology; and (iii) the presence of spillovers meaning that second movers enjoy the same increased return to the cleaner technology than the early investors.

The temporal dimension introduced with the above conditions leads to nontrivial strategic effects. Non-signatories play a game of their own, where the maximal safe investment in the 'dirty' technology is determined by the number of those that showed leadership by restricting themselves in its use. In particular, the interplay of (i)–(iii) can either catalyse or deter investments in the clean technology, by affecting participation in the treaty and, consequently, the incentives for the fringe. From the point of view of the fringe, the presence of leaders (i) has potentially conflicting effects. This is because of the coexistence of increased free riding incentives (the target is within reach and second movers may optimistically assume that others will take it upon themselves to restrict their use of the polluting technology) and opposite incentives to cooperate (early commitments to the common good may entice the fringe to follow

suit). The increased competitiveness of the socially preferable technology (ii) will affect the two groups differently depending on whether there are spillovers to the fringe (iii).

Perhaps unsurprisingly, the two treatments where subjects cooperated the most are those in which it is less costly to do so—that is, when the gap between the returns of the clean and polluting investments is smallest. The distribution of burdens between the two groups, however, is rather different. While most of it is undertaken by the coalition when its members retain the benefits of cooperation on innovation, the reverse is true when these benefits trickle through to the fringe: in that case, coalition size drops by about 40%, but the fringe, lured by the spillovers, embraces the new technology. This effect is even more marked when the magnitude of the innovation benefits is smaller. In this case, the drop in coalition size is more modest under positive spillovers, while the effect on fringe behaviour persists, which leads to a significantly higher chance of avoidance of crossing the threshold for dangerous climate change.

These findings point to the importance of finding a mechanism to change players' incentives in a threshold public bad game and to create reinforcement effects between members' and non-members' strategies. In order to achieve a common goal, such as climate change mitigation to stabilize emissions at a safe level, which is beyond reach by a subgroup acting alone and requires widespread cooperation, leaders have to find a way to involve the outsiders. In climate negotiations, one possible approach is adding R&D and innovation to the bargaining table. Reducing the cost-effectiveness gap with respect to the incumbent technology (e.g., fossil fuels) by means of investments by a set of motivated innovators, may lure more reluctant players toward an environmentally superior but individually costlier alternative. Our results suggest that all parties may benefit when followers can also profit from these investments, especially when investments in R&D can only provide limited returns. The challenge is striking a balance by designing a mechanism that involves outsiders without deterring (too much) cooperation from the leaders.

Of course, caution must be used when extrapolating to climate negotiations. The problem faced by real negotiators has many more layers of complexity that will make the matter of coordination more difficult. Moreover, implementing agreements, such as the one negotiated in Paris and recently entered into force, that rely on nationally determined pledges, is likely be further hindered by myopic policymaking (Barrett and Dannenberg 2016). In the context of this

game, that would mean that actual investments in the 'dirty' technology would be higher than those agreed to by signatories.

We maintain that the possibility to signal willingness to be part of a coalition of climate leaders is not enough, and it is therefore vital to induce the participation of the more reluctant players. Our experimental findings suggest that the diffusion of innovation may be an important lever for climate action.

Supplementary material

Supplementary material is available online at the OUP website. This includes the Appendices as well as the raw experimental data, the instructions for other treatments, the comprehension questionnaire and the Ztree codes.

Funding

This work was supported by IEFE Bocconi and the European Research Council grant [ERC-2013-StG 336703-RISICO to V.B.]. M.H. acknowledges funding from the Marie Curie fellowship INTCOP. A.T. acknowledges the financial support of the Grantham Foundation for the Protection of the Environment as well as that of the ESRC Centre for Climate Change Economics and Policy, which is funded by the United Kingdom's Economic and Social Research Council.

Acknowledgements

The authors are grateful to Jérémy Celse for his invaluable help. We also thank Andreas Lange for his comments on a previous draft, as well as Davide Rossi, Laura Dell'Acqua, and Michele Peruzzi for their assistance in running the experiments. In addition, we are grateful to the seminar participants at ASFEE, WCERE 2014, ISEE 2014, to the organizers and participants of the Bath and FEEM workshops, as well as to the referees and editor of the Journal for their constructive comments. Logistic support from the Bocconi Experimental Laboratory in the Social Sciences (BELSS), which hosted our experimental sessions, is kindly acknowledged.

References

Alley, R. B., J. Marotzke, W. D. Nordhaus, J. T. Overpeck, D. M. Peteet, R. A. Pielke, R. T. Pierrehumbert, et al. 2003. 'Abrupt Climate Change.' *Science* 299: 2005–2010.

Andreoni, J. 1995. 'Warm-Glow versus Cold-Prickle: The Effects of Positive and Negative Framing on Cooperation in Experiments.' *The Quarterly Journal of Economics*, 1–21.

Bagnoli, M. and M. McKee. 1991. 'Voluntary Contribution Games: Efficient Private Provision of Public Goods.' *Economic Inquiry* 29: 351–366.

Barrett, S. 1994. 'Self-Enforcing International Environmental Agreements.' *Oxford Economic Papers*, 878–894.

Barrett, S. 2003. 'Environment and Statecraft: The Strategy of Environmental Treaty-Making.' *Management of Environmental Quality: An International Journal* 14: 622–623.

Barrett, S. 2006. 'Climate Treaties And' breakthrough' technologies.' *The American Economic Review* 96: 22–25.

Barrett, S. and A. Dannenberg. 2012. 'Climate Negotiations under Scientific Uncertainty.' *Proceedings of the National Academy of Sciences* 109: 17372–17376.

Barrett, S., and A. Dannenberg. 2016. 'An experimental investigation into 'pledge and review' in climate negotiations.' *Climatic Change* 138: 339-351.

Burger, N. E., and C. D. Kolstad. 2009. 'Voluntary Public Goods Provision, Coalition Formation, and Uncertainty.' National Bureau of Economic Research. http://www.nber.org/papers/w15543.

Carraro, C. and D. Siniscalco. 1993. 'Strategies for the International Protection of the Environment.' *Journal of Public Economics* 52: 309–328.

Carraro, C. and D. Siniscalco. 1995. 'R&D Cooperation and the Stability of International Environmental Agreements.' http://ftp.cepr.org/active/publications/discussion_papers/dp.php?dpno=1154. [Accessed on December 29, 2016]

Cherry, T. L. and D. M. McEvoy. 2013. Enforcing compliance with environmental agreements in the absence of strong institutions: An experimental analysis. *Environmental and Resource Economics*, 54: 63-77.

Clark, P. 2016. ' EU t o fast-track P aris climate change agreement'. *Financial Times*. Available at https://www.ft.com/content/edb484f8-8716-11e6-a75a-0c4dce033ade [Accessed on October 3, 2016]

Dannenberg, A., A. Lange, and B. Sturm. 2014. 'Participation and Commitment in Voluntary Coalitions to Provide Public Goods.' *Economica* 81: 257–275.

Dannenberg, A., A. Löschel, G. Paolacci, C. Reif, and A. Tavoni. 2015. 'On the Provision of Public Goods with Probabilistic and Ambiguous Thresholds.' *Environmental and Resource Economics* 61: 365–383.

Dechezleprêtre, A., M. Glachant, and Y. Ménière. 2008. 'The Clean Development Mechanism and the International Diffusion of Technologies: An Empirical Study.' *Energy Policy* 36: 1273–1283.

Di Maria, C. and E. Van der Werf. 2008. 'Carbon Leakage Revisited: Unilateral Climate Policy with Directed Technical Change.' *Environmental and Resource Economics* 39: 55–74.

Eaton, J. and S. Kortum. 2001. 'Trade in Capital Goods.' *European Economic Review* 45: 1195–1235.

Eaton, J. and S. Kortum. 2006. 'Innovation, Diffusion, and Trade.' National Bureau of Economic Research. http://www.nber.org/papers/w12385. [Accessed on December 29, 2016]
Edenhofer, O., R. P. Madruga, Y. Sokona, E. Farahani, S. Kadner, K. Seyboth, A. Adler, et al. 2014. 'Summary for Policymakers Climate Change 2014, Mitigation of Climate Change.' *IPCC 2014, Climate Change 2014: Contribution of Working Group III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*.
Finus, M. 2001. *Game Theory And International Environmental Cooperation*. Edward Elgar, Cheltenman.
Fischbacher, U. 2007. 'Z-Tree: Zurich Toolbox for Ready-Made Economic Experiments.' *Experimental Economics* 10: 171–78.
Golombek, R. and M. Hoel. 2004. 'Unilateral Emission Reductions and Cross-Country Technology Spillovers.' *Advances in Economic Analysis & Policy* 3.
Granovetter, M. 1978. 'Threshold Models of Collective Behavior.' *American Journal of Sociology*, 1420–1443.
Heal, G. and H. Kunreuther. 2011. 'Tipping Climate Negotiations.' National Bureau of Economic Research Working Paper No. 16954.
Heugues, M. 2014. 'International Environmental Cooperation: A New Eye on the Greenhouse Gas Emissions' Control.' *Annals of Operations Research* 220: 239–262.
İriş, D., J. Lee, and A. Tavoni. 2016. 'Delegation and Public Pressure in a Threshold Public Goods Game: Theory and Experimental Evidence.' Centre for Climate Change Economics and Policy Working Paper No. 211. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2752895. [Accessed on December 29, 2016]
Keller, W. 2010. 'International Trade, Foreign Direct Investment, and Technology Spillovers.' *Handbook of the Economics of Innovation* 2: 793–829.
Kolstad, C., K. Urama, J. Broome, A. Bruvoll, M. C. Olvera, D. Fullerton, C. Gollier, et al. 2014. 'Social, Economic and Ethical Concepts and Methods.' In *Climate Change 2014: Mitigation of Climate Change. Contribution of Working Group III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change [Edenhofer, O., R. Pichs-Madruga, Y. Sokona, E. Farahani, S. Kadner, K. Seyboth, A. Adler, I. Baum, S. Brunner, P. Eickemeier, B. Kriemann, J. Savolainen, S. Schlömer, C. von Stechow, T. Zwickel and J.C. Minx (Eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.*
Kosfeld, M., A. Okada, and A. Riedl. 2009. 'Institution Formation in Public Goods Games.' *The American Economic Review* 99: 1335–1355.
Kriegler, E., J. W. Hall, H. Held, R. Dawson, and H. J. Schellnhuber. 2009. 'Imprecise Probability Assessment of Tipping Points in the Climate System.' *Proceedings of the National Academy of Sciences* 106: 5041–5046.
Lade, S. J., A. Tavoni, S. A. Levin, and M. Schlüter. 2013. 'Regime Shifts in a Social-Ecological System.' *Theoretical Ecology* 6: 359–372.
Lenton, T. M., H. Held, E. Kriegler, J. W. Hall, W. Lucht, S. Rahmstorf, and H. J. Schellnhuber. 2008. 'Tipping Elements in the Earth's Climate System.' *Proceedings of the National Academy of Sciences* 105: 1786–1793.
Levati, M. V., M. Sutter, and E. Van der Heijden. 2007. 'Leading by Example in a Public Goods Experiment with Heterogeneity and Incomplete Information.' *Journal of Conflict Resolution* 51: 793–818.

Liberman, V., S. M. Samuels, and L. Ross. 2004. 'The Name of the Game: Predictive Power of Reputations versus Situational Labels in Determining Prisoner's Dilemma Game Moves.' *Personality and Social Psychology Bulletin* 30: 1175–1185.

Marchiori, C., S. Dietz, and A. Tavoni. 2017. 'Domestic Politics and the Formation of International Environmental Agreements.' *Journal of Environmental Economics and Management* 81: 115-131.

McEvoy, D. M., J. J. Murphy, J. M. Spraggon, and J. K. Stranlund. 2011. 'The problem of maintaining compliance within stable coalitions: experimental evidence.' *Oxford Economic Papers* 63:475-498.

McEvoy, D. M., T. L. Cherry, and J. K. Stranlund. 2015. 'Endogenous Minimum Participation in International Environmental Agreements: An Experimental Analysis.' *Environmental and Resource Economics* 62: 729–744.

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

Moxnes, E. and E. Van der Heijden. 2003. 'The Effect of Leadership in a Public Bad Experiment.' *Journal of Conflict Resolution* 47: 773–795.

Ostrom, E. 2009. 'A Polycentric Approach for Coping with Climate Change.' *Available at SSRN 1934353*. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1934353.

Rockström, J., W. Steffen, K. Noone, A. Persson, F. S. Chapin, E. F. Lambin, T. M. Lenton, et al. 2009. 'A Safe Operating Space for Humanity.' *Nature* 461: 472–475.

Seres, S., E. Haites, and K. Murphy. 2009. 'Analysis of Technology Transfer in CDM Projects: An Update.' *Energy Policy* 37: 4919–4926.

Sonnemans, J., A. Schram, and T. Offerman. 1998. 'Public Good Provision and Public Bad Prevention: The Effect of Framing.' *Journal of Economic Behavior & Organization* 34: 143–161.

Sterner, T. and M. Damon. 2011. 'Green Growth in the Post-Copenhagen Climate.' *Energy Policy* 39: 7165–7173.

Tavoni, A. 2013. 'Game Theory: Building up Cooperation.' *Nature Climate Change* 3: 782–783.

Tavoni, A., A. Dannenberg, G. Kallis, and A. Löschel. 2011. 'Inequality, Communication, and the Avoidance of Disastrous Climate Change in a Public Goods Game.' *Proceedings of the National Academy of Sciences* 108: 11825–11829.

Vasconcelos, V. V., F. C. Santos, and J. M. Pacheco. 2013. 'A Bottom-up Institutional Approach to Cooperative Governance of Risky Commons.' *Nature Climate Change* 3: 797–801.

Watts, D. J. 2002. 'A Simple Model of Global Cascades on Random Networks.' *Proceedings of the National Academy of Sciences* 99: 5766–5771.

Weir, S., and J. Knight. 2004. 'Externality Effects of Education: Dynamics of the Adoption and Diffusion of an Innovation in Rural Ethiopia.' *Economic Development and Cultural Change* 53: 93–113.