Centre for Climate Change Economics and Policy

An ESRC Research Centre

Grantham Research Institute on Climate Change and the Environment

Tipping points and loss aversion in international environmental agreements

Doruk İriş and Alessandro Tavoni

May 2016

Centre for Climate Change Economics and Policy
Working Paper No. 269

Grantham Research Institute on Climate Change and the Environment
Working Paper No. 239

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

The Grantham Foundation for the Protection of the Environment

UNIVERSITY OF LEEDS

LSE The London School of Economics and Political Science

The Centre for Climate Change Economics and Policy (CCCEP) was established by the University of Leeds and the London School of Economics and Political Science in 2008 to advance public and private action on climate change through innovative, rigorous research. The Centre is funded by the UK Economic and Social Research Council. Its second phase started in 2013 and there are five integrated research themes:

1. Understanding green growth and climate-compatible development
2. Advancing climate finance and investment
3. Evaluating the performance of climate policies
4. Managing climate risks and uncertainties and strengthening climate services
5. Enabling rapid transitions in mitigation and adaptation

More information about the Centre for Climate Change Economics and Policy can be found at: http://www.cccep.ac.uk.

The Grantham Research Institute on Climate Change and the Environment was established by the London School of Economics and Political Science in 2008 to bring together international expertise on economics, finance, geography, the environment, international development and political economy to create a world-leading centre for policy-relevant research and training. The Institute is funded by the Grantham Foundation for the Protection of the Environment and the Global Green Growth Institute. It has nine research programmes:

1. Adaptation and development
2. Carbon trading and finance
3. Ecosystems, resources and the natural environment
4. Energy, technology and trade
5. Future generations and social justice
6. Growth and the economy
7. International environmental negotiations
8. Modelling and decision making
9. Private sector adaptation, risk and insurance

More information about the Grantham Research Institute on Climate Change and the Environment can be found at: http://www.lse.ac.uk/grantham.

Tipping Points and Loss Aversion in International Environmental Agreements

Doruk İriş[1], Alessandro Tavoni[2]

Abstract

We study the impact of loss-aversion and the threat of catastrophic damages, which we jointly call threshold concerns, on international environmental agreements. We aim to understand whether a threshold for dangerous climate change is an effective coordination device for countries to overcome the free-riding problem, so that they abate emissions sufficiently to avoid disaster. We focus on loss-averse countries negotiating under the threat of either high environmental damages (loss domain) or low damages (gain domain). Under symmetry, when countries display identical degrees of threshold concern, we show that such beliefs have a positive effect on reducing the emission levels of both signatories to the treaty and non-signatories, leading to weakly larger coalitions of signatories. We then introduce asymmetry, by allowing countries to differ in the degree of concern about the threat of disaster. We show that stable coalitions are mostly formed by the countries with higher threshold concerns. When enough countries have no threshold concern, coalition size may diminish, regardless of whether the other countries have mild or strong threshold concerns.

[1]School of Economics, Sogang University, Mapo-gu, Seoul, 04107, S. Korea.
Phone: +82-2-705-8505, Email: dorukiris@sogang.ac.kr
[2]Grantham Research Institute on Climate Change and the Environment, London School of Economics, London WC2A 2AZ, UK. Tavoni is supported by the Centre for Climate Change Economics and Policy, which is funded by the UK Economic and Social Research Council (ESRC).
The authors would like to thank Frank Venmans and participants of SED2015.

1. Introduction

The theory of international environmental agreements (IEAs) has produced stark insights into the difficulties of achieving cooperation. Due to the intrinsic trade-off between the breadth of the agreement, as measured by the number of acceding countries, and the depth of the abatement commitments, game theorists have postulated that self-enforcing environmental agreements will have limited success. Either few signatories will commit to stringent targets, or many countries will sign on to a shallow agreement that only achieves modest reductions (Barrett, 1994; Carraro and Siniscalco, 1993; d'Aspremont et al., 1983; Hoel, 1992). The standard model has recently been extended to account for important empirical findings, including: introducing asymmetric countries and the possibility of making side payments, relaxing rationality and perfect foresight assumptions ascribed to countries, and linkage of cooperation on IEAs with other issues such as trade and R&D (for reviews of this literature, see Barrett, 2005, and Finus, 2008). One feature, which is common to virtually all IEA literature, is that reference considerations are absent from countries' welfare functions. These depend on absolute benefits and costs of emissions in a continuous fashion.

In economics and psychology, the concept of loss aversion has recently been used to account for the empirical finding that individuals place a higher weight on losses than gains, violating the assumption of standard economic theory that tastes are unchanging (Kahneman, 2003). Theories of loss aversion have sprung up with proposed explanations for this ubiquitous phenomenon, occurring in financial markets (e.g., Benartzi & Thaler, 1995), consumption and savings patterns (e.g., Bowman et al., 1999), macroeconomic policy (e.g., Ciccarone & Marchetti, 2013), contract theory (e.g., Daido et al., 2013), real estate transactions (e.g., Genesove & Mayer, 2001), the energy paradox (Greene, 2011), competitive behavior (e.g., Eisenkopf & Teyssier, 2013), and trade (Freund & Ozden, 2008; Tovar, 2009), among others.

Remarkably, loss aversion has not, to the best of our knowledge, been used in modeling environmental agreements.[1] Given the pervasiveness of reference point considerations in human decision-making, we investigate its role in affecting the size and commitment level of coalitions cooperating on curbing emission levels in the presence of loss aversion with respect to a threshold amount for acceptable environmental damage. The premise is that there exists a "tipping point," which is viewed by all states as indicative of an approaching catastrophe (Tavoni and Levin, 2014). That is, nations believe that below a given tolerable amount of environmental damage business carries on as usual, according to the standard calculus of net benefits from pollution, but above a critical level of damage from emissions, additional losses ensue according to a multiplier effect.

The literature on environmental tipping points and disastrous climate change has recently focused on such boundary conditions, which, if crossed, may trigger quick and unavoidable ecosystem collapse (Scheffer et al., 2001; Lade et al., 2013). Rockström and colleagues (Rockström et al., 2009) identified planetary boundaries that define "the safe operating space for humanity with respect to the Earth system and are associated with the planet's biophysical subsystems or processes." They suggest that the boundaries in three systems, including climate change (for which they propose to keep atmospheric carbon dioxide concentration below 350 parts per million and the change in radiative forcing below one watt per square meter), have already been crossed. Hence, the prospect of incurring additional losses from ecosystem collapse may well enter into governments' considerations. This will be particularly likely for vulnerable developing countries with limited capability to adapt to the changing climate, for instance those that are located on coastal areas and are prone to flooding.

[1] One exception is İriş (forthcoming). It examines the implications of political parties being averse to insufficient economic performance (relative to a critical economic target level) on sustaining an international environmental agreement in an infinitely repeated game setting. Other widely used behavioral concepts that have been incorporated into IEAs are reciprocity (Hadjiyiannis et al., 2012; Nyborg, 2015) and inequity-aversion (Lange, 2006). See İriş and Tavoni (2016), which reviews the literature on tipping points and reference-dependent preferences in climate change games.

In this paper we concentrate on the implications (in terms of stability and breadth of a stylized IEA) of enriching the standard model by introducing countries' aversion to environmental losses, together with a concern for exceeding a so-called tipping point, i.e., a critical level of admissible damages beyond which disastrous consequences are expected. We refer to these preferences as threshold concerns, and note that one can recover the standard model without loss aversion by setting one parameter equal to zero, as discussed on page eight.

For tractability reasons, in Section 2, we abstract from such asymmetries in exposure to the damages arising from high concentrations of pollutants, and assume that countries are symmetric and uniformly perceive the threshold for catastrophic damages, given by T. Introducing uncertainty on the location of the threshold can destabilize coordination by reverting the game to a prisoner's dilemma (Barrett, 2013).

The related experimental literature on the provision of discrete public goods subject to thresholds corroborates this result.[2] It has been shown that both asymmetries among players, as well as uncertainty about the location of the threshold hinder group achievement as measured by the likelihood of avoidance of the dangerous equilibrium where catastrophic losses occur (Tavoni et al., 2011; Dannenberg et al., 2015). On the other hand, leadership appears to be an important engine of collective action, as successful experimental groups tend to eliminate inequality over the course of the game. In these, rich players signal willingness to redistribute their funds early on in the game (Tavoni et al., 2011). Related studies confirm the importance of leadership (Bosetti et al., 2015; Dietz et al., 2012; İriş et al., 2015), especially on the part of wealthy actors (Vasconcelos et al., 2014).

We can thus view the theoretical model presented in Section 2 as an initial step in introducing realistic features in the standard coalition formation model of international environmental agreements. We anticipate that the symmetry and common knowledge assumptions utilized in Section 2 are likely to bias upwards the

[2] For a more detailed review on this experimental literature, see İriş and Tavoni (2016).

transformative potential of the threshold in fostering cooperation. We check for this effect in Section 3, which extends the model by introducing some degree of asymmetry in countries' threshold concerns. More specifically, we extend the model to allow countries to have different beliefs on the environmentally tolerable level of pollution, by letting a fraction of the countries believe that the critical threshold is higher than the one perceived by the remaining countries.[3]

In ecological processes, threshold uncertainty is often irreducible; nevertheless, scientists often attach probabilities to different future environmental scenarios. For example, the 2013 Intergovernmental Panel on Climate Change's Summary for Policymakers (IPCC, 2013) states that: "There is high confidence that sustained warming greater than some threshold would lead to the near-complete loss of the Greenland ice sheet over a millennium or more, causing a global mean sea level rise of up to 7 m. Current estimates indicate that the threshold is greater than about 1°C (low confidence) but less than about 4°C (medium confidence) global mean warming with respect to pre-industrial." Hence, early warning signals, if picked up and correctly processed in time, may act as stimuli for action on environmental protection.

We also investigate theoretically this hypothesis by introducing aversion to losses in excess of the given threshold T, which can be viewed as reflecting the scientific or political consensus on what level of environmental damage is deemed tolerable. In the case of climate change, where unsafe levels of warming (e.g., 4°C) have been linked to damages (e.g., loss of the Greenland ice sheet), one can also interpret T in terms of temperature change generally associated with catastrophic climate change. That is, levels of warming beyond which environmental damages increase abruptly and are subject to irreversibility. As mentioned above, we do away with the

[3] Section 3 of the current paper contributes to the literature, which study the implications of country asymmetries on IEAs. Kolstad (2010) examines countries' asymmetries in their size and marginal damage from pollution; McGinty (2007) and Pavlova and de Zeeuw (2013) in their marginal costs and benefits of abatement; and Mendez and Trelles (2000) in their technologies.

complexities arising from uncertainty over the threshold level. Under this optimistic scenario where no uncertainty muddles the value of the safe pollution level, we ask whether the traditionally negative prediction of either small or ineffective international environmental agreements can be reverted (Barrett, 1994; Carraro and Siniscalco, 1993).

Under symmetric threshold concerns, we show that the form of loss-aversion we used has a positive effect on reducing the emission levels of both signatories and non-signatories, leading to a larger coalition in some cases. Therefore, countries are more likely to take on significant environmental commitments when they believe they face the threat of an impending environmental catastrophe.

Under asymmetric threshold concerns, stable coalitions are mostly formed by the countries with higher threshold concerns. The size of the coalition diminishes when enough countries lack a concern for overstepping the threshold, regardless of the preferences of the other countries (whether they have strong or mild threshold concerns). Unlike in the symmetric setup, where the stable coalitions are always unique, under asymmetry, uniqueness is not guaranteed: in some cases, a coalition may not form; in others, more than one stable coalition can materialize.

Our model closely follows and extends Diamantoudi and Sartzetakis (2006); DS, henceforth. In Section 2, we introduce the basic notions of the model under the assumption that countries are symmetric. In Subsection 2.1, we study two benchmark cases, the games associated with non-cooperative behavior, and full cooperative behavior. In Subsection 2.2, we introduce the coalition formation game, which consists of non-signatory behavior, signatory behavior, and the stability analysis (to determine the size of the stable IEA). In Section 3, we extend the model by allowing different countries to have differing degrees of aversion to environmental losses.

2. Symmetric Model

We consider a regional or global pollution game involving n identical countries, $N = \{1, 2, \dots, n\}$. Production and consumption in each country i generates emissions e_i of a transnational pollutant. Pollution is a public bad, that is, each country's emission not only damages itself, but also damages other countries in equal measure, thus imposing a negative externality on others. We assume that each country i simultaneously decides its non-negative emission level, $e_i \geq 0$.[4] By this assumption, we exclude the possibility of an existing stock of pollution that can be diminished through abatement efforts. The standard social welfare of country i is the difference between i's benefits from emissions $B_i(e_i)$ due to production and consumption and the transboundary environmental damages $D_i(E)$ from the aggregate emissions, $E = \sum_i^n e_i$. We use the following quadratic functional forms for the benefit and damage functions:

$$B_i(e_i) = \beta e_i - \frac{1}{2} e_i^2, \quad \text{and} \quad D_i(E) = \frac{\gamma}{2} E^2, \tag{1}$$

where β and γ are positive.

In addition to the standard calculus outlined above, we assume that each country i has concerns on the level of environmental damages and whether it exceeds a critical threshold $T \geq 0$ representing the environmentally safe operating limit.[5] If the level of environmental damages does not exceed the threshold, i.e., $D_i(E) \leq T$, then each country i enjoys being in safe territory. If the level of environmental damages exceeds the threshold, $D_i(E) > T$, then each country's welfare drops due to the threat of an environmental catastrophe. Specifically, we assume that governments are averse to environmental losses, i.e., they have a stronger tendency to avoid the environmental losses generated by large emissions than acquiring gains

[4] Instead of emissions, abatement effort could be used as the choice variable; see, for instance, Barrett (1994). DS show that the two choices are strategically equivalent.

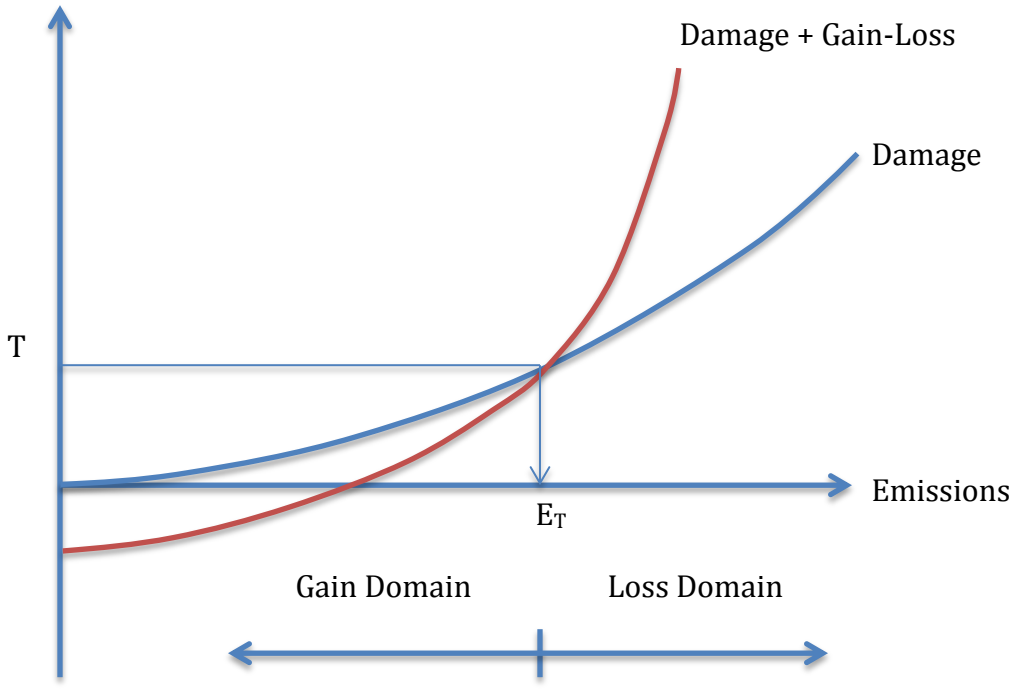
[5] The extension in Section 3 captures countries' asymmetry on the perception of the environmentally safe operating limit.

(through increased emissions). The environmental gain-loss function of country i is written as follows:

$$GL_i(E,T) = \begin{cases} T - D_i(E), & D_i(E) \leq T \\ \lambda\big(T - D_i(E)\big), & D_i(E) > T \end{cases} \quad (2)$$

for $\lambda > 1$, where λ is known as a loss-aversion parameter.[6]

Figure 1: Environmental Damage and Gain-Loss Functions.



Note: the loss-averse countries' damage function is steeper in all domains and much steeper in the loss domain, compared to the gain domain. This is due to the kink caused by the loss-aversion parameter λ.

The social welfare of loss-averse country i depends on its own emissions as well as on the emissions of others, $e_{-i} = \{e_1, \dots, e_{i-1}, e_{i+1}, \dots, e_n\}$, in addition to depending on the threshold for the environmentally safe operating limit T:

$$w_i(e_i, e_{-i}, T) = \beta e_i - \frac{1}{2}e_i^2 - \frac{\gamma}{2}(E)^2 + \alpha \begin{cases} T - \frac{\gamma}{2}(E)^2, & \frac{\gamma}{2}(E)^2 \leq T \\ \lambda\left(T - \frac{\gamma}{2}(E)^2\right), & \frac{\gamma}{2}(E)^2 > T \end{cases} \quad (3)$$

[6] This well-known formulation is a local definition of loss aversion by Köbberling and Wakker (2005)

where α is a positive scaling factor, determining the degree to which country i cares about the environmental gain-loss function.

A loss-averse country incorporates the gain-loss function to the damage function and, thus, to the social welfare function. As shown in Figure 1, the adjusted environmental damage is steeper over the entire domain, compared to the case without the gain-loss function. It always incentivizes countries to lower their emissions. However, it is much steeper in the loss domain than in the gain domain, owing to the kink caused by the loss-aversion parameter $\lambda > 1$ at the threshold.

For convenience, we focus our analysis on the more general case where the social welfare of a loss-averse country i can be written as follows:

$$w_i(e_i, E, T) = \beta e_i - \frac{1}{2}e_i^2 - \frac{\gamma}{2}LE^2 + (L-1)T \qquad (4)$$

where L captures threshold concerns. L incorporates both the degree to which governments care about the environmental gain-loss function (α) and the level of aversion to environmental losses (λ), whenever applicable. It takes different values in the following three possible cases:

$$L(E) = \begin{cases} 1, \ no\ threshold\ \ (neutral\ domain) \\ 1+\alpha, \ D(E) \leq T\ \ (gain\ domain) \\ 1+\alpha\lambda, \ D(E) > T\ (loss\ domain) \end{cases} . \qquad (5)$$

If the level of environmental damages exceeds T, substituting $L = (1 + \alpha\lambda)$ in (4) results in the loss domain. The gain domain results instead when $\lambda = 1$, implying $L = (1 + \alpha)$. Similarly, the neutral domain is recovered by equating $\alpha = 0$, which implies $L = 1$.

We further assume that a loss-averse country maximizes social welfare in (4) as if it were in the gain domain ($L = 1 + \alpha$). If environmental damages are indeed lower than the threshold, $D(E) \leq T$, then its emissions and social welfare are determined. However, if environmental damages turn out to be higher than the threshold, then the country is in the loss domain, and $L = (1 + \alpha\lambda)$ is instead used in the maximization of social welfare (4). For some parameter values, country i may not exceed the threshold anymore when $L = (1 + \alpha\lambda)$ is employed. However, we

assume that it will not switch back to using $L = 1 + \alpha$, since it would lead to a loss for the country.

Note that this analysis for a general L is possible since the critical threshold T disappears once the first-order condition is taken. Thus, once the domain is determined, the threshold only levies the social welfare level but not the chosen emission levels.

2.1. Two Benchmark Cases: The Non-cooperative and Full Cooperation Cases

The non-cooperative case relies on the standard Cournot/Nash equilibrium in which countries pursue their unilateral strategies. Given the emission levels of the other countries, each country chooses its emission level to maximize the social welfare function described in (4). In order to derive the equilibrium emission level, first, we find the best-response function by taking the first-order condition of the maximization problem and equating it to zero, $\partial w_i(.)/\partial e_i = 0$,[7]

$$e_i\left(\sum_{j\neq i} e_j\right) = \frac{\beta - \gamma L\left(\sum_{j\neq i} e_j\right)}{1+\gamma L}. \qquad (6)$$

Under symmetry, all countries emit the same in equilibrium. Substituting the emission level of all countries by the non-cooperative emission level, e_{nc}, yields the non-cooperative equilibrium emission level: $e_{nc} = \frac{\beta}{1+n\gamma L}$. Observe that the non-cooperative emissions decrease in countries' threshold concerns L. Substituting e_{nc} into (4) gives the non-cooperative welfare:

$$w_{nc} = \frac{\beta^2\left(1-n\gamma L(n-2)\right)}{2(1+n\gamma L)^2} + (L-1)T. \qquad (7)$$

In the full cooperation case, all countries choose how much to jointly emit to maximize their aggregate social welfare function, $w = \sum_{i=1}^{n} w_i$. The solution of the maximization problem is found by $\partial w(.)/\partial e_i = 0$. The full cooperative outcome yields the following per-country emission level, $e_c = \frac{\beta}{1+n^2\gamma L}$. Cooperative emissions

[7] To increase readability, we avoid a significant amount of simple but tedious calculations in the paper. Nevertheless, we can provide a Mathematica supporting file for these calculations upon request, either in pdf or nb format.

decrease in countries' threshold also concerns L. Substituting e_c into (4) gives the cooperative welfare:

$$w_{nc} = \frac{\beta^2}{2(1+n^2\gamma L)} + (L-1)T. \quad (8)$$

While both non-cooperative and cooperative emission levels decrease in countries' threshold concerns, as expected by embedding the gain-loss function into the social welfare, the drop in emissions does not necessarily imply an increase in welfare levels. Welfare in both the non-cooperative and cooperative solutions consists of two counteracting parts. The first terms in (7) and (8) decrease in L due to the amplified perceived damages, while the second terms increase in L due to the stronger weight placed on the threshold.[8]

2.2. Partial Cooperation

The coalition formation game consists of three stages that are solved simultaneously, assuming that countries can look forward and infer backwards. Stage 1 is a participation game in which each country chooses simultaneously to be either a signatory or a non-signatory to a stylized IEA. Stages 2 and 3 entail a Stackelberg game with signatories playing the role of leaders. More specifically, the signatories jointly decide their emission levels in Stage 2, followed by non-signatory countries independently deciding their emission levels in Stage 3. The game is solved using backward induction.

A set of countries $S \subset N$ signs an agreement, while the remaining N\S countries do not. The coalition, formed by $|S| = s$ signatories, generates emissions E_s, with each member emitting e_s such that $E_s = se_s$. Each non-signatory emits e_{ns}, so that non-signatories collectively emit $E_{ns} = (n-s)e_{ns}$.

The non-signatory countries are thus the Stackelberg followers, i.e., they observe the actions of the signatories, and then act non-cooperatively given the emission level of the leaders and other non-signatory countries. The behavior of non-signatories is

[8] This tradeoff also materializes in the partial cooperation setting, and when countries have asymmetric threshold concerns. We thus omit similar welfare analyses for those cases.

described by the same best-response function as in the non-cooperative model (5). Since, by symmetry, all non-signatory countries emit the same level in equilibrium, e_{ns}, the other countries, except non-signatory i, emit jointly $(n-s-1)e_{ns}+se_s$, yielding the best-response function depending on signatories' emission level:

$$e_{ns}(e_s)=\frac{\beta-\gamma Lse_s}{X} \tag{9}$$

where $X=1+\gamma L(n-s)$. Signatories are the Stackelberg leaders, i.e., they know how the non-signatory countries best respond to their emission levels, and so they take it into account and act cooperatively with the other signatory countries. More formally, they maximize the objective function, $w^S=\sum_{i\in S}w_i$, by solving $\partial w^S(.)/\partial e_s=0$, subject to the best response function $e_{ns}(e_s)$ in (8). The emission level of a signatory is,

$$e_s=\beta\left(1-\frac{\gamma Lns}{\Psi}\right), \tag{10}$$

where $\Psi=\gamma s^2L+X^2$. Substituting signatory countries' emission level (9) into the non-signatory's best response function (8) gives the emission level of a non-signatory:

$$e_{ns}=\beta\left(1-\frac{\gamma LnX}{\Psi}\right)=e_s+\frac{\beta\gamma Ln(s-X)}{\Psi}. \tag{11}$$

Note that $s>X$ should hold for $e_{ns}>e_s$, which is equivalent to $\gamma<\frac{s-1}{(n-s)L}$. Moreover, the aggregate emission level of all countries simplifies to the following:

$$E=E_s+E_{ns}=se_s+(n-s)e_{ns}=\frac{\beta nX}{\Psi}. \tag{12}$$

We need to guarantee the signatory and non-signatory countries' emission levels to be positive, which is satisfied by the conditions below:[9]

$$e_s>0\Rightarrow\gamma<\frac{4}{n\,L\,(n-4)}\text{ for }n>4;\ e_{ns}>0\Rightarrow\gamma<\frac{4}{n\,L\,(n-4)}\text{ for }n\ >\ 4.$$

These conditions require the relative impact of damages to benefits to be not very high. Having non-trivial threshold concerns (that is, departing from the standard

[9] The proof of this condition and all other proofs are in the appendix.

model of loss neutrality, with $L > 1$) additionally requires the relative impact of damages to be smaller. DS find a very similar condition without the threshold concerns L. As they point out, this apparently harmless condition is essential and restricts the size of the stable coalition to be 2, 3, or 4.

Next, we obtain the indirect social welfare functions of signatory countries, ω_s, and non-signatory countries, ω_{ns}, by substituting the relevant emission levels of the signatories and the non-signatories and aggregate emissions (10-12) into the social welfare function:

$$\omega_s = \beta^2\left(\frac{1}{2} - \frac{\gamma L n^2}{2\Psi}\right) + (L-1)T, \text{ and } \omega_{ns} = \beta^2\left(\frac{1}{2} - \frac{\gamma L(1+\gamma L)n^2X^2}{2\Psi^2}\right) + (L-1)T. \quad (13)$$

The following Lemma, similar to proposition 2 in DS, defines the properties of indirect welfare functions.

Lemma 1: Consider the indirect welfare functions of signatory and non-signatory countries, ω_s and ω_{ns}, respectively, and let $z^{min} = \frac{1+\gamma L n}{1+\gamma L}$. Then,

i. $z^{min} = argmin_{s\in\mathbb{R}\cap[0,n]}\omega_s$;

ii. $\omega_s(s)$ increases in s if $s > z^{min}$ and it decreases in s if $s < z^{min}$;

iii. $\omega_{ns}(s) > (<)\omega_s(s)$ for all $s > (<)z^{min}$.

iv. If, moreover, z^{min} is an integer, then the two indirect welfare levels are equal at $s = z^{min}$, that is, $\omega_{ns}\left(z^{min}\right) = \omega_s\left(z^{min}\right)$.

Lemma 1 shows that a country is better off as a signatory when the size of the coalition is small, and that its welfare decreases as the size of the coalition increases. Next, we discuss the impact of governments' threshold concerns on the welfare functions.

Proposition 1: Let $L'' > L'$, then

i. $z^{min}(L'') > z^{min}(L')$ for $n > 1$.

ii. For all $\tilde{s} \in \left(z^{min}(L'), z^{min}(L'')\right)$, $\omega_s(s,L)|_{s=\tilde{s},L=L'}$ increases in s and $\omega_s(s,L)|_{s=\tilde{s},L=L''}$ decreases in s. For any other $s \notin \left(z^{min}(L'), z^{min}(L'')\right)$, if $\omega_s(s,L)|_{L=L'}$ decreases (increases), $\omega_s(s,L)|_{L=L''}$ decreases (increases).

iii. For all $\tilde{s} \in \left(z^{min}(L'), z^{min}(L'')\right)$, $\omega_{ns}(s,L)|_{s=\tilde{s},L=L'} > \omega_s(s,L)|_{s=\tilde{s},L=L'}$ and $\omega_{ns}(s,L)|_{s=\tilde{s},L=L''} < \omega_s(s,L)|_{s=\tilde{s},L=L''}$.

The main finding of Proposition 1 is that there are some coalition sizes such that a country would be better off as a non-signatory when the threshold concerns are relatively low. However, for the same coalition sizes, a country would be better off as a signatory when countries' threshold concerns are relatively high.

2.2.1. Stable Coalition

We have already found the emission levels of signatory and non-signatory countries in Stages 2 and 3. We now solve the participation game in Stage 1, to determine the number of signatories s^* in a stable coalition. A coalition is stable if it satisfies internal and external stability conditions, which guarantee that the agreement is self-enforcing. The conditions are, respectively:

$$\omega_s(s^*) \geq \omega_{ns}(s^* - 1) \text{ and } \quad \omega_s(s^* + 1) \leq \omega_{ns}(s^*). \tag{14}$$

The internal stability condition guarantees that a signatory country cannot be better off by unilaterally leaving the coalition. Similarly, the external stability condition guarantees that a non-signatory country cannot be better off by unilaterally joining the coalition.[10]

The existence and uniqueness of a stable coalition for the social welfare functions with the additional gain-loss function follows DS's Proposition 3. More specifically, as DS show, for $n > 4$, there exists a unique stable coalition whose size is $s^* \in \{2,3,4\}$. Next, we analyze how a change in countries' threshold concerns affects the stable coalition size.

[10] The conditions (14) are first used for cartel stability by d'Aspremont et al. (1983), then adapted to international public goods cooperation by Barrett (1992, 1994), Hoel (1992), and Carraro and Siniscalco (1993).

Proposition 2: For $n > 4$, $\partial s^*/\partial L \geq 0$.

We are going to illustrate the findings of Proposition 2 with a numerical example in which the size of the stable coalition increases from 2 to 3. In this example, we assume $n = 10$, $\beta = 5/3$, $\gamma = 0.01$, and $L(= 1 + \alpha\lambda) \leq 1.5$, which guarantees the condition for positive emissions to hold: If $\gamma < \frac{4}{n\,L\,(n-4)} \Leftrightarrow 0.01 < 0.04\overline{4}$.

Figure 2 depicts the case when governments do not exhibit concerns for dangerous climate change beyond a tipping point, $L = 1$. Figure 3 focuses instead on countries with some degree of threshold concern: we set $L = 1.5$ for visual clarity.[11] While T does not play any role in Figure 2, T is set to be 1 in Figure 3, which places countries in the loss domain.[12] In both figures, the indirect welfare function $\omega_s(s)$ is represented by the solid curve, $\omega_{ns}(s)$ by the dotted curve, and $\omega_{ns}(s-1)$ by the dashed curve. All the indirect welfare functions are depicted against the size of coalitions s, and here the range is restricted to the values of interest, $s = 1, \dots, 4$.

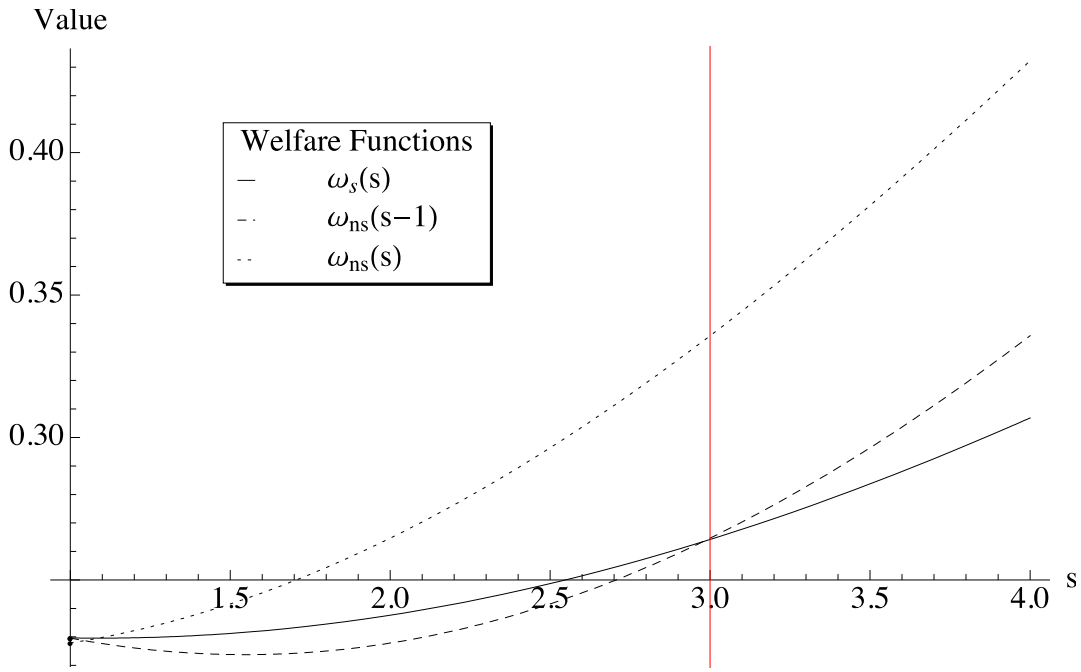
In Figure 2, one can observe that coalition size $s^* = 2$ is internally stable, $\omega_s(s^*) \geq \omega_{ns}(s^* - 1)$, since the solid curve is above the dashed curve at $s = 2$. Note also that these two curves intersect at $s = 2.976$, so $s = 3$ is not internally stable. Moreover, coalition size $s^* = 2$ is also externally stable, $\omega_s(s^* + 1) \leq \omega_{ns}(s^*)$, since the dotted curve is above the dashed curve at $s = 3$. Therefore, the coalition size $s^* = 2$ is stable.

In Figure 3, one can follow similar arguments and observe that coalition size $s^* = 3$ is both internally and externally stable. Therefore, the stable coalition size weakly increases as threshold concerns are introduced (or concerns become stronger), when the environmentally safe operating limits are exceeded.

[11] In this numerical example, it is sufficient to set $L \geq 1.02551$ for the coalition size to increase from 2 to 3.

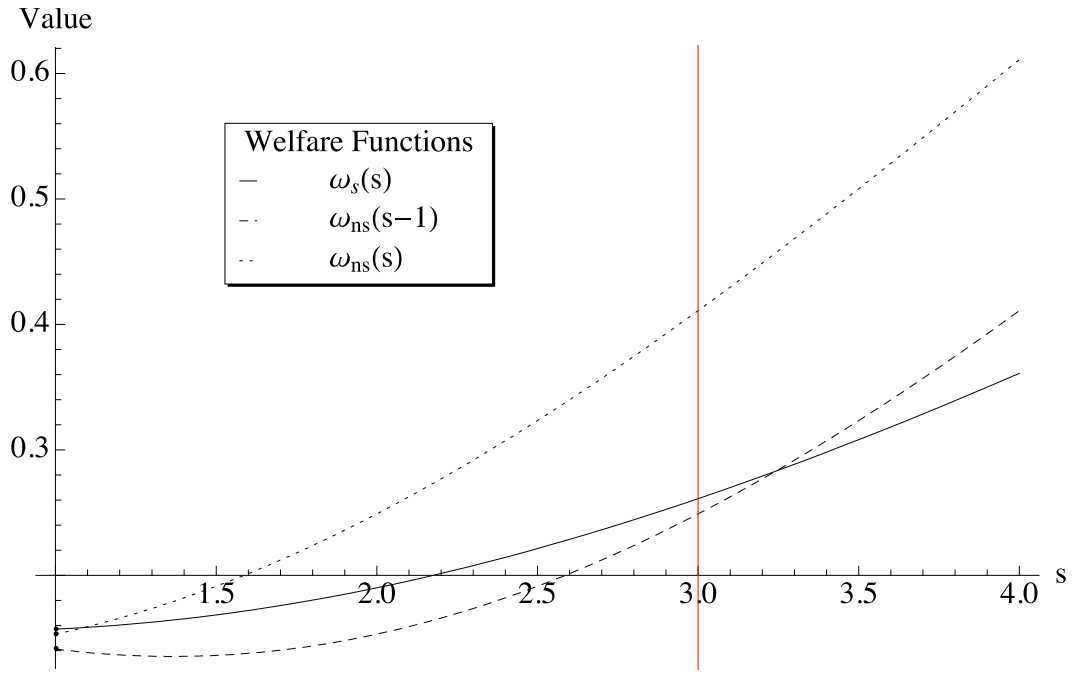
[12] Remember that T does not affect the emission levels, once the domain is determined. It does levy the welfare level, but in equal measure for all welfare functions $\omega_s(s)$, $\omega_{ns}(s)$, and, $\omega_{ns}(s-1)$. Thus, the size of the coalition does not depend on T so long as countries remain in the same domain (gain, loss, or neutrality).

Figure 2: Coalition Size without Threshold Concerns ($L = 1$)
$n = 10, \beta = 5/3, \gamma = 0.01$.



Note: The stable coalition size is $s^* = 2$

Figure 3: Coalition Size with Threshold Concerns ($L = 1.5$)
$n = 10, \beta = 5/3, \gamma = 0.01$.



Note: The stable coalition size is $s^* = 3$

3. Asymmetric Model

In this section, we consider the case when, out of n countries, h have a high concern for exceeding the threshold and $n-h$ have low threshold concerns: $L_h = 1 + \alpha_h\lambda_h > L_l = 1 + \alpha_l\lambda_l$. Alternatively, one can interpret this as h countries having low and $n-h$ countries having high environmentally safe operating limits $T_l \leq T_h$. Thus, h countries are in the loss domain, and $n-h$ countries are either in the gain or neutral domain.

3.1. Two Benchmark Cases: The Non-cooperative and Full Cooperation Cases

Similar to the symmetric case, in the non-cooperative case countries maximize their welfare, according to (4). However, the problem for country i differs depending on the degree of concern, as follows:

$$w_{hi}(e_{hi}, E, T) = \beta e_{hi} - \frac{1}{2}e_{hi}^2 - \frac{\gamma}{2}\, L_h(e_{hi} + (h-1)e_h + (n-h)e_l\,)^2 + \,(L_h - 1)T;$$

$$w_{li}(e_{li}, E, T) = \beta e_{li} - \tfrac{1}{2}e_{li}^2 - \tfrac{\gamma}{2}\, L_h(e_{li} + he_h + (n-h-1)e_l\,)^2 + \,(L_l - 1)T; \quad (15)$$

where e_{hi} and e_{li} are the emission levels of country i, and e_h and e_l are any other country's emission levels with high and low threshold concerns. The FOCs, $\partial w_{hi}(.)/\partial e_{hi} = 0$ and $\partial w_{li}(.)/\partial e_{li} = 0$, give the best-response functions for a country with high and low threshold concerns, respectively:

$$e_{hi}\big((h-1)e_h + (n-h)e_l\big) = \tfrac{\beta - \gamma L_h\left((h-1)e_h + (n-h)e_l\,\right)}{1+\gamma L_h};$$

$$e_{li}(he_h + (n-h-1)e_l) = \tfrac{\beta - \gamma L_l(he_h + (n-h-1)e_l\,)}{1+\gamma L_l}. \quad (16)$$

In equilibrium, countries with the same level of threshold concerns emit the same, that is $e_{hi} = e_h$ and $e_{li} = e_l$. Thus, the best-response functions for any country with high and low threshold concerns are:

$$e_h\big((n-h)e_l\big) = \tfrac{\beta - \gamma L_h(n-h)e_l}{1+\gamma h L_h} \qquad \text{and} \qquad e_l(he_h) = \tfrac{\beta - \gamma L_l h e_h}{1+\gamma L_l(n-h)}. \quad (17)$$

Substituting one into the other gives the non-cooperative equilibrium emissions:

$$e_h^{nc} = \beta\left(\frac{1-\gamma(n-h)(L_h-L_l)}{1+\gamma(hL_h+(n-h)L_l)}\right) \quad \text{and} \quad e_l^{nc} = \beta\left(\frac{1+\gamma h(L_h-L_l)}{1+\gamma(hL_h+(n-h)L_l)}\right). \tag{18}$$

Notice that the denominators of both emission levels are the same. Then, it is straightforward to observe that countries with high threshold concerns emit less than the ones with low threshold concerns in the non-cooperative solution: $e_h^{nc} < e_l^{nc}$. Furthermore, $\gamma < \frac{1}{(L_h-L_l)(n-h)}$ should hold for $e_h^{nc} > 0$.

In the full cooperation case, both types of countries jointly decide their emission levels to maximize their aggregate social welfare function, $w = \sum_{i=1}^{n} w_i$. The solution of the maximization problem is found by setting $\partial w(.)/\partial e_h = 0$ and $\partial w(.)/\partial e_l = 0$, and substituting one into the other. This yields the same emission levels for both types of countries:

$$e_h^c = e_l^c = \frac{\beta}{1+\gamma n(hL_h+(n-h)L_l)}. \tag{19}$$

3.2. Partial Cooperation

We are now going to study a similar coalition formation game to the one in section 2.2, by solving the asymmetric participation game so as to derive the number of signatories. Both countries with high and low threshold concerns can now be signatories to the treaty, and we denote them respectively by s_h and s_l, with $s = s_h + s_l$. That means the numbers of non-signatories with high and low threshold concerns are respectively $h - s_h$ and $n - h - s_l$. The best-response function, governing the behavior of non-signatory country i with high threshold concerns, is written as follows:

$$e_{nshi}(e_{sh}, e_{sl}, e_{nsh}, e_{nsl}) = \frac{\beta-\gamma L_h\left((n-h-s_l)e_{nsl}+(h-s_h-1)e_{nsh}+s_h e_{sh}+s_l e_{sl}\right)}{1+\gamma L_h} \tag{20}$$

where e_{sh}, e_{sl}, e_{nsh} and e_{nsl} are emission levels of signatory and non-signatory countries with high (L_h) and low (L_l) threshold concerns, respectively. Since all non-signatory countries with high threshold concerns have the best-response function in (18), we set $e_{nshi} = e_{nsh}$ and find their best-response functions:

$$e_{nsh}(e_{sh}, e_{sl}, e_{nsl}) = \frac{\beta-\gamma L_h\left((n-h-s_l)e_{nsl}+s_h e_{sh}+s_l e_{sl}\right)}{1+\gamma L_h(h-s_h)}. \tag{21}$$

One can follow the same steps for non-signatories with low threshold concerns, and find the following best-response function:

$$e_{nsl}(e_{sh}, e_{sl}, e_{nsh}) = \frac{\beta - \gamma L_l\left((h - s_h)e_{nsh} + s_h e_{sh} + s_l e_{sl}\right)}{1 + \gamma L_l(n - h - s_l)}. \tag{22}$$

Since all non-signatories simultaneously decide their emission levels after observing the emission levels of the signatories, we substitute one into the other and find the best-response functions for non-signatories with high and low threshold concerns, depending on the emissions of the signatories only:

$$e_{nsh}(e_{sh}, e_{sl}) = \frac{\beta - \gamma L_h(s_h e_{sh} + s_l e_{sl}) - \beta\gamma(n - h - s_l)(L_h - L_l)}{Y} \tag{23}$$

$$e_{nsl}(e_{sh}, e_{sl}) = \frac{\beta - \gamma L_l(s_h e_{sh} + s_l e_{sl}) - \beta\gamma(h - s_h)(L_h - L_l)}{Y}$$

where $Y = 1 + \gamma\big(L_h(h - s_h) + L_l(n - h - s_l)\big)$.

Signatories maximize their joint welfare function, $w^S = \sum_{i \in S} w_i$, which consists of signatory countries with both high and low threshold concerns, subject to the best-response functions of non-signatories in (23). Integrating these best-response functions into the joint welfare and solving the problem by the FOCs ($\partial w^S(.)/\partial e_{sh} = 0$ and $\partial w^S(.)/\partial e_{sl} = 0$) yields the emission levels of both types of signatories. These depend on each other's emission level, as follows:

$$e_{sh}(e_{sl}) = \frac{\beta Y^2 - \gamma(\beta(n - s_h - s_l) + s_l e_{sl})(s_h L_h + s_l L_l)}{Y^2 + \gamma s_h(s_h L_h + s_l L_l)}; \tag{24}$$

$$e_{sl}(e_{sh}) = \frac{\beta Y^2 - \gamma(\beta(n - s_h - s_l) + s_h e_{sh})(s_h L_h + s_l L_l)}{Y^2 + \gamma s_h(s_h L_h + s_l L_l)}.$$

Since signatories decide their emission level simultaneously, we substitute one into the other, which gives the emission levels of signatory countries with high and low threshold concerns.

$$e_{sh} = e_{sl} = \frac{\beta\left(\Omega - \gamma n(s_h L_h + s_l L_l)\right)}{\Omega} \tag{25}$$

where $\Omega = Y^2 + \gamma(s_h + s_l)(s_h L_h + s_l L_l)$. Substituting (25) back into the non-signatory countries' best-response functions in (23) gives the emission level of non-signatory countries with high and low threshold concerns, respectively:

$$e_{nsh} = \frac{\beta(\Omega - \gamma L_h n Y)}{\Omega} \quad \text{and} \quad e_{nsl} = \frac{\beta(\Omega - \gamma L_l n Y)}{\Omega}. \tag{26}$$

Next, we find the aggregate emission level of all countries under the asymmetric case $E^A = E_{sh} + E_{sl} + E_{nsh} + E_{nsl}$, which simplifies to the following expression:

$$E^A = s_h e_{sh} + s_l e_{sl} + (h - s_h) e_{nsh} + (n - h - s_l) e_{nsl} = \frac{\beta n Y}{\Omega}. \tag{27}$$

Note that $e_{nsh} < e_{nsl}$, as expected. For $e_{sh} = e_{sl} < e_{nsh}$, we need $\gamma < \frac{(s_h - 1)L_h + s_l L_l}{(h - s_h)L_h + (n - h - s_l)L_h L_l}$. Furthermore, the condition $\gamma < \frac{1}{(s_h L_h + s_l L_l)(n - h - s_l)}$ suffices for $e_{sh} = e_{sl} > 0$ to hold. As in the symmetric case, all of these conditions require the relative impact of damages to benefits to not be very high.

Lastly, we obtain the indirect social welfare functions of signatories and non-signatories with high and low threshold concerns by substituting the relevant emission levels from (25-26) into the social welfare function:

$$\begin{aligned}
\omega_{sh} &= \beta^2 \left(\frac{1}{2} - \frac{\gamma n^2 \left(Y^2 L_h + \gamma (s_h L_h + s_l L_l) \right)}{2\Omega^2} \right) + (L_h - 1)T; \\
\omega_{sl} &= \beta^2 \left(\frac{1}{2} \frac{\gamma n^2 \left(Y^2 L_l + \gamma (s_h L_h + s_l L_l) \right)}{2\Omega^2} \right) + (L_l - 1)T; \\
\omega_{nsh} &= \beta^2 \left(\frac{1}{2} - \frac{\gamma n^2 Y^2 L_h (1 + \gamma L_h)}{2\Omega^2} \right) + (L_h - 1)T; \\
\omega_{nsl} &= \beta^2 \left(\frac{1}{2} - \frac{\gamma n^2 Y^2 L_l (1 + \gamma L_l)}{2\Omega^2} \right) + (L_l - 1)T.
\end{aligned} \tag{28}$$

3.3. Stability Analysis

In our asymmetric model, a coalition is stable if it satisfies internal and external stability conditions for countries with high and low threshold concerns:

$$\omega_{sh}(s_h^*, s_l^*, h, n) \geq \omega_{nsh}(s_h^* - 1, s_l^*, h, n), \omega_{sh}(s_h^* + 1, s_l^*, h, n) \leq \omega_{nsh}(s_h^*, s_l^*, h, n); \tag{29}$$

$$\omega_{sl}(s_h^*, s_l^*, h, n) \geq \omega_{nsl}(s_h^*, s_l^* - 1, h, n), \omega_{sl}(s_h^*, s_l^* + 1, h, n) \leq \omega_{nsl}(s_h^*, s_l^*, h, n). \tag{30}$$

Due to the asymmetry, these conditions depend on the number of signatories of either kind: s_h^* and s_l^*. This requires all four conditions to be satisfied. For instance, given a number of signatory countries with low threshold concerns s_l^-, the stable

number of countries with high threshold can be s_h^-. However, given s_h^-, s_l^- might not be a stable number of countries with low threshold concerns. Moreover, these conditions also depend on the number of countries with high (h) and low ($n-h$) threshold concerns. Varying h changes these conditions and what types of countries form a stable coalition, as we show below.

In the following three tables, we present the results of our numerical analysis on the stable number of signatories with different levels of threshold concerns. In each table, the four rows show the number signatory countries with low threshold concerns $s_l \in \{0,1,2,3\}$. Similarly, the columns show the number signatory countries with high threshold concerns $s_h \in \{0,1,2,3\}$. Columns are grouped by different number of countries with high threshold concerns $h \in \{0,1,\dots,10\}$. The unfeasible columns are omitted, since for any h, we have $s_h \leq h$.

For each column, s_h equals 0, 1, 2, or 3; the conditions in (30) provide a stable number of signatories with low threshold concerns s_l^*, and we mark the respective cell with "l." Similarly, for each row, s_l equals 0, 1, 2, or 3; the conditions in (29) provide a stable number of signatories with high threshold concerns s_h^*, and we mark the respective cell with "h." If one cell contains both "h" and "l," then it shows how many signatories with high and low threshold concerns form this stable coalition.

In this numerical example, we assume $n = 10, \beta = 5/3,$ and $\gamma = 0.03333333332$. The conditions on positive emissions and signatories emitting less than non-signatories are satisfied, i.e., $0 < e_{sh} = e_{sl} < e_{nsh} < e_{nsl}$ for all scenarios described below.

Table 1: Stable Number of Signatories with High ($L_h = 2$) and Low ($L_l = 1.5$) Threshold Concerns

	h=0	h=1		h=2			h=3				h=4,5				h=6				h≥7			
sl\sh	0	0	1	0	1	2	0	1	2	3	0	1	2	3	0	1	2	3	0	1	2	3
0						l			l	h			l	h		l	l	h	l	l	l	h
1						h			h			l	h				h				h	
2			h		h			h				h			l	h				h		
3	l	l	h	l	h		l	h			l	h				h				h		

Note: shaded areas indicate the stable coalitions and the number of signatories with high and low threshold concerns

In Table 1, we assume $L_h = 2$ and $L_l = 1.5$. This is a scenario in which both types of countries have significant threshold concerns but one group has stronger concerns than the other. Several interesting findings are worth noting. First, for any h, the size of the stable coalitions is $s_h^* + s_l^* = 3$. Second, for $h \geq 4$, the stable coalition only consists of countries with high threshold concerns, $(s_h^*, s_l^*) = (3, 0)$. Third, for $h = 3$, two stable coalitions exist, $(s_h^*, s_l^*) \in \{(3,0), (1,2)\}$. Fourth, for $h \in \{1, 2, 3\}$, two countries with low threshold concerns sign up to a stable coalition.

Table 2: Stable Number of Signatories with High $(L_h = 2)$ and No $(L_l = 1)$ Threshold Concerns

	h=0	h=1		h=2,3				h≥4			
sl\sh	0	0	1	0	1	2	3	0	1	2	3
0			l		l	h			l	l	h
1						h				h	
2	l					h				h	
3		l	h	l		h		l		h	

Note: shaded areas indicate the stable coalitions and the number of signatories with high and low threshold concerns

In Table 2 we assume $L_h = 2$ and $L_l = 1$. This is a scenario in which one type of country has significant threshold concerns, but the other has none. Compared to the case presented in Table 1, the asymmetry between these two types of countries is much more severe, leading to the following findings. First, for $h \leq 3$, the size of stable coalitions $s_h^* + s_l^* = 2$. Second, countries with low threshold concerns have weaker incentives to participate in any coalition, due to stronger external effects. Countries with high threshold concerns have stronger incentives to participate for $h \geq 4$, and also if some countries with low threshold concerns participate. However, for $h \leq 3$ and $s_l = 0$, they have weaker incentives as well. Third, observe that a stable coalition may not exist.

Table 3: Stable Number of Signatories with Mild ($L_h = 1.1$) and No ($L_l = 1$) Threshold Concerns

	h=0	h=1		h=2,3				h=4,5,6,7				h≥8			
sl\sh	0	0	1	0	1	2	3	0	1	2	3	0	1	2	3
0						h	l			h	l			l	h
1			l		l	h			l	h			l	h	
2	l	l	h	l	h				h				h		
3		h		h				h				h			

Note: shaded areas indicate the stable coalitions and the number of signatories with high and low threshold concerns

In Table 3, we assume that $L_h = 1.1$ and $L_l = 1$. This is a scenario in which one type of country has mild threshold concerns, but the other has none. Note also that this case has the weakest asymmetry between two types of countries, leading to the following findings. First, compared to the case presented by Table 2, countries with no threshold concerns ($L_l = 1$) have stronger incentives to participate, because weaker asymmetry between types implies weaker external effects. Second, observe again the multiplicity and potential non-existence of stable coalitions. We observe the multiplicity of stable coalitions even if there is an equal number of countries with high and low threshold concerns, $h = n - h = 5$.

In sum, we observe that countries with higher threshold concerns tend to form most of the coalitions. However, countries with low threshold concerns may also join the coalition if they are relatively high in number, i.e., for low h's. One type of country having no threshold concern could cause the coalition size to diminish, regardless of the other countries having strong or mild threshold concerns. This can be also due to the decrease in countries' aggregate threshold concerns. Finally, a unique stable coalition always exists under symmetry. However, stable coalitions may not exist, or more than one stable coalition can exist once asymmetry in the threshold concerns is introduced.

4. Discussion

We have studied the impact of loss-aversion and reference dependence on the breadth and stability of an international environmental agreement aimed at abating emissions in the presence of the threat of dangerous climate change. We model it as a perceived tipping point, a threshold level of damages from emissions of pollutants linked with industrial production, beyond which severe losses may be incurred. In the symmetric case, which allows for greater analytical traction, we assume that every country shares the same views on the entity of the threshold. Hence, heterogeneity arises only with respect to the number of countries signing up to an IEA in this setting. We then extend the model to allow for the more realistic case where countries differ in their beliefs about the threshold for dangerous climate change. Such differing views may originate from uncertainty about the location of the threshold for dangerous climate change, or from the difficulty in translating a given threshold into the effort required to avoid overstepping such a boundary, as argued in Barrett (2013).

We have shown that loss aversion reduces global emission levels relative to the standard model, even though it does not necessarily increase countries' welfare, either under full cooperation or when countries act non-cooperatively. We have further established that, under some conditions, loss aversion has a similar effect on the emission levels of both signatories and non-signatories to an IEA, potentially leading to a larger coalition size. We conclude that loss-averse countries are more likely to take significant environmental decisions on reducing their emissions, provided that their governments perceive that there is a credible threat of an approaching environmental catastrophe.

The degree of variation among the beliefs held by different countries negotiating climate change abatement is of course an empirical matter. Here we abstract from real world subtleties and assume, for the sake of tractability, either symmetric behavior or a minimalistic level of heterogeneity with either high or low level of concern for the environmental losses. Introducing asymmetric perceptions on the

presence and location of the tipping point (ideally backed by empirical evidence), appears to be a fruitful avenue of extension of the stylized model we have introduced here. This appears to be particularly salient at the moment, given that a significant part of the discussion in the 2015 climate summit in Paris revolved around whether countries should collectively aim for a 1.5°C or 2°C increase in average global temperature.

Recent literature has developed to analyze the effect of tipping points on climate change cooperation, some of which we have briefly reviewed here. We have added to it by introducing a related behavioral aspect, loss aversion, a pervasive trait among humans. Loss aversion is particularly salient for problems such as climate change, which largely pertain to the loss domain, especially when contemplating the damages arising from dangerous climate change. We hope that the simple model we have introduced here will stimulate further research on this topic, which is interestingly located at the nexus of economics, behavior, and ecology.

References:

Barrett, S., 1992. International environmental agreements as games. In: Pethig, R. (Ed.), Conflict and Cooperation in Managing Environmental Resources. Springer-Verlag, Berlin, pp. 11–37.

Barrett, S., 1994. Self-enforcing international environmental agreements. Oxford Economic Papers 46, 878–894.

Barrett, S., 2005. The theory of international environmental agreements. In: Maeler, K.-G., Vincent, J. (Eds.), Handbook of Environmental Economics. Vol. 3. Elsevier, Amsterdam, pp. 1457–1516.

Barrett, S., 2013. Climate Treaties and Approaching Catastrophes. Journal of Environmental Economics and Management, 66 (2): 235-250.

Benartzi, S., Thaler, R. H., 1995. Myopic loss aversion and the equity premium puzzle. The Quarterly Journal of Economics 110 (1), 73-92.

Bosetti, V., Heugues, M., Tavoni, A. Luring others in: Coalition formation games with threshold and spillover effects. Centre for Climate Change Economics and Policy Working Paper No. 199.

Bowman, D., Minehart, D., Rabin, M., 1999. Loss aversion in a consumption–savings model. Journal of Economic Behavior & Organization 38 (2), 155-178.

Carraro, C., Siniscalco, D., 1993. Strategies for international protection of the environment. Journal of Public Economics 52, 309–328.

Ciccarone, G., Marchetti, E., 2013. Rational expectations and loss aversion: Potential output and welfare implications. Journal of Economic Behavior & Organization 86, 24-36.

d'Aspremont, C., Jacquemin, A., Gabszewicz, J., Weymark, J., 1983. On the stability of collusive price leadership. Canadian Journal of Economics 16 (1), 17–25.

Daido, K., Morita, K., Murooka, T., Ogawa, H., 2013. Task assignment under agent loss aversion. Economics Letters 121 (1), 35-38.

Dannenberg, A., Löschel, G., Paolacci, C., Reif, A., Tavoni A., 2015. On the Provision of Public Goods with Probabilistic and Ambiguous Thresholds. Environmental and Resource Economics 61 (3), 365-383.

Dietz, S., Marchiori, C., Tavoni, A., 2012. Why do we see unilateral action on climate change? Vox 5.

Diamantoudi, E., Sartzetakis, E., 2006. Stable international environmental agreements: an analytical approach. Journal of Public Economic Theory 8 (2), 247-263.

Eisenkopf, G., Teyssier, S., 2013. Envy and loss aversion in tournaments. Journal of Economic Psychology 34, 240-255.

Finus, M., 2008. Game theoretic research on the design of international environmental agreements: insights, critical remarks, and future challenges. International Review of Environmental and Resource Economics 2 (1), 29–67.

Freund, C, Özden Ç., 2008. Trade policy and loss aversion. The American Economic Review 98 (4), 1675-1691.

Genesove, D, Mayer, C, 2001. Loss aversion and seller behavior: Evidence from the housing market. The Quarterly Journal of Economics 116 (4), 1233-1260.

Greene, D. L., 2011. Uncertainty, loss aversion, and markets for energy efficiency. Energy Economics 33 (4), 608-616.

Hadjiyiannis, C., İriş, D., and Tabakis, C. (2012). International environmental cooperation under fairness and reciprocity. The B.E. Journal of Economic Analysis & Policy, 12(1):1–30.

Hoel, M., 1992. International environment conventions: the case of uniform reductions of emissions. Environmental and Resource Economics 2 (2), 141–159.

İriş, D. (forthcoming). Economic Targets and Loss-Aversion in International Environmental Cooperation. Journal of Economic Surveys, Special Issue: Economics of Climate Change.

İriş, D., Lee, J., Tavoni, A, 2015. 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.

İriş, D., Tavoni, A, (forthcoming). Tipping and reference points in climate change games. Handbook on the Economics of Climate Change, Chichilnisky and Rezai (eds.), Edward Elgar Press, UK.

IPCC, Summary for Policymakers, in Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, T. F. Stocker et al., Eds. Cambridge Univ. Press, Cambridge UK.

Kahneman, D., 2003. A psychological perspective on economics. The American Economic Review 93 (2), 162-168.

Kolstad, C. D., 2010. Equity, Heterogeneity and International Environmental Agreements. The B.E. Journal of Economic Analysis & Policy, 10(2):1–17.

Köbberling, V., Wakker, P. P., 2005. An index of loss aversion. Journal of Economic Theory 122 (1), 119-131.

Lade, S., Tavoni, A., Levin, S., Schlüter, M., 2013. Regime shifts in a social-ecological system. Theoretical Ecology 6, 359-372.

Lange, A., 2006. The Impact of Equity-Preferences on the Stability of International Environmental Agreements. Environmental and Resource Economics, 34, 247-267.

McGinty, M., 2007. International environmental agreements among asymmetric nations. Oxford Economic Papers, 59(1):45–62.

Mendez, L., Trelles, R., 2000. The abatement market a proposal for environmental cooperation among asymmetric countries. Environmental and Resource Economics, 16(1):15–30.

Nyborg, K., 2015. Reciprocal Climate Negotiators. IZA Discussion Paper No. 8866.

Pavlova, Y., de Zeeuw, A., 2013. Asymmetries in international environmental agreements. Environment and Development Economics, 18:51–68.

Rockström J, et al., 2009. A safe operating space for humanity. Nature 461, 472–475.

Scheffer, M., Carpenter, S., Foley, J. A., Folke, C., Walker, B., 2001. Catastrophic shifts in ecosystems. Nature 413, 591-596.

Tavoni, A., Dannenberg, A., Kallis, G., Löschel, A., 2011. Inequality, communication and the avoidance of disastrous climate change in a public goods game. Proceedings of the National Academy of Sciences 108 (29), 11825–11829.

Tavoni, A., Levin, S., 2014. Managing the Climate Commons at the Nexus of Ecology, Behaviour and Economics. Nature Climate Change 4, 1057-1063.

Tovar, P., 2009. The effects of loss aversion on trade policy: Theory and evidence. Journal of International Economics 78 (1), 154-167.

Vasconcelos, V.V., F.C. Santos, J.M. Pacheco, and S.A. Levin, 2014. Climate policies under wealth inequality. Proc Natl Acad Sci USA, 111(6): p. 2212-2216.

Appendix:

The number of signatories s is a non-negative integer smaller than the number of countries. In the proofs, we treat s as a real number in $[0, n]$ and convert it to an integer at the end whenever necessary.

Proof of the Condition for positive emissions:

From equation (7), we have $e_s = \beta\left(1 - \frac{\gamma Lns}{\gamma s^2 L + X^2}\right)$. For $e_s > 0$, the following condition should hold: $1 + \gamma L(n-s)(\gamma L(n-s) - (s-2)) > 0$. Let $A(s) = 1 + \gamma L(n-s)(\gamma L(n-s) - (s-2))$ and $\underline{s} = argmin_s\, A(s) = \frac{2+\mathrm{n}+2\gamma \mathrm{Ln}}{2(1+\gamma \mathrm{L})}$. For $A(s) > 0$ for any s, it is sufficient to show that $A(\underline{s}) > 0$. One can easily find that $A(\underline{s}) = \frac{4-\gamma \mathrm{L}(\mathrm{n}-4)\mathrm{n}}{4(1+\gamma L)}$ and for $A(\underline{s}) > 0$, we need $\gamma < \frac{4}{n\,L\,(n-4)}$.

From equation (8), we have $e_{ns} = \beta\left(1 - \frac{\gamma LnX}{\Psi}\right)$. For $e_{ns} > 0$, the following condition should hold: $(1 + \gamma L(n-s))(1 - \gamma Ls) + \gamma Ls^2 > 0$. Let $\Phi(s) = (1 + \gamma L(n-s))(1 - \gamma Ls) + \gamma Ls^2$ and $\underline{s} = argmin_s\, \Phi(s) = \frac{2+\gamma Ln}{2(1+\gamma L)}$. For $\Phi(s) > 0$ for any s, it is sufficient to show that $\Phi(\overline{s}) > 0$. One can easily find that $\Phi(\overline{s}) = \left(\frac{2+\gamma \mathrm{Ln}(2+\gamma \mathrm{L})}{2(1+\gamma L)}\right)\left(\frac{2-\gamma^2 L^2 n}{2(1+\gamma L)}\right) + \left(\frac{\gamma L(2+\gamma Ln)^2}{4(1+\gamma L)^2}\right)$ and for $\Phi(\overline{s}) > 0$, it is sufficient to have $\frac{2-\gamma^2 L^2 n}{2(1+\gamma L)} > 0 \Leftrightarrow \gamma < \frac{1}{L}\sqrt{\frac{2}{n}}$. Note that $\frac{4}{n\,L\,(n-4)} < \frac{1}{L}\sqrt{\frac{2}{n}}$ for $n \geq 6$. At $n = 5$, we have $\Phi(\overline{s}) = \frac{4+20\gamma \mathrm{L}-25\gamma^3 \mathrm{L}^3}{4(1+\gamma L)}$ and for $\Phi(\overline{s}) > 0$, the following condition should hold: $25\gamma^3 \mathrm{L}^3 - 20\gamma \mathrm{L} - 4 < 0$ and indeed holds for $\gamma < \frac{4}{5L}$. QED.

Proof of Lemma 1:

i. Let us first find z^{min} by taking the partial derivative of the signatory welfare function with respect to the number of signatories and equate it to zero, which will simplify to the following:

$$\frac{\partial \omega_s}{\partial s} = \frac{(\beta\gamma Ln)^2(s-X)}{\Psi^2} = 0.$$

For the equality to hold, we need $s = X$, thus, $s = 1 + \gamma L(n - s)$. Solving for s gives, $s = z^{min} = \frac{1+\gamma Ln}{1+\gamma L}$. Since $\frac{\partial^2 \omega_s}{\partial s^2} > 0$ for all β, γ, n, the FOC is sufficient.

ii. Observe that $\frac{\partial \omega_s}{\partial s} > (<) 0$ if $s > (<) X \Leftrightarrow s > (<) z^{min}$.

iii. Using the indirect welfare functions, we can write ω_{ns} in terms of ω_s:

$\omega_{ns} = \omega_s + \frac{(\beta\gamma Ln)^2(s-X)(s+X)}{2\Psi^2}$.

It is straightforward to observe that $\omega_{ns} \lesseqgtr \omega_s$, for $s \lesseqgtr X \Leftrightarrow s \lesseqgtr z^{min}$.

iv. Finally, if z^{min} is an integer, then for $s = z^{min} \Leftrightarrow s = X$ and $\omega_{ns}\left(z^{min}\right) = \omega_s\left(z^{min}\right)$.

Proof of Proposition 1:

i. $\frac{\partial z^{min}}{\partial L} = \frac{\gamma n(1+\gamma L) - \gamma(1+\gamma Ln)}{(1+\gamma L)^2} = \frac{\gamma(n-1)}{(1+\gamma L)^2} > 0$ for n>1.

ii. By the second bullet of Lemma 1, for any $\tilde{s} \in \left(z^{min}(L'), z^{min}(L'')\right)$, $\omega_s(s, L)|_{s=\tilde{s}, L=L'}$ increases in s since $\tilde{s} > z^{min}(L')$. Similarly, the second bullet of Lemma 1 implies that $\omega_s(s, L)|_{s=\tilde{s}, L=L''}$ decreases in s since $\tilde{s} < z^{min}(L'')$.

For any other $s \notin \left(z^{min}(L'), z^{min}(L'')\right)$, a higher environmental threshold concern does affect how the number of signatories changes the welfare of signatories. Thus, if $\omega_s(s, L)|_{L=L'}$ decreases (increases), $\omega_s(s, L)|_{L=L''}$ decreases (increases) as well.

iii. The third bullet of Lemma 1 implies that for all $\tilde{s} \in \left(z^{min}(L'), z^{min}(L'')\right)$, $\omega_{ns}(s, L)|_{s=\tilde{s}, L=L'} > \omega_s(s, L)|_{s=\tilde{s}, L=L'}$, and $\omega_{ns}(s, L)|_{s=\tilde{s}, L=L''} < \omega_s(s, L)|_{s=\tilde{s}, L=L''}$.

Proof of Proposition 2: Remember that $\omega_s\left(z^{min}\right) = \omega_{ns}\left(z^{min}\right)$. Let us define $\bar{z} = z^{min} + 1$ and let z' be the smallest s such that $\omega_s(z') = \omega_{ns}(z' - 1)$. DS show, in the proof of Proposition 3, that $\bar{z} < z' < \bar{z} + 1$. Moreover, DS prove that if $z' < 3$,

then $s^* = 2$, if $z' < 4$ then $s^* = 3$, and if $z' \geq 4$, then $s^* = 4$. By the definition of $\bar{z}$, we can write the condition as $z^{min} + 1 < z' < z^{min} + 2$. It is then straightforward to observe that for an increase in L, which increases z^{min}, the size of the stable coalition would weakly increase. QED.