Loss Aversion in International Environmental Agreements

Doruk İriş[1], Alessandro Tavoni[2]

Abstract

We study the impact of loss-aversion and the threat of critical damages from insufficient pollutant abatement, which we jointly call threshold concerns, on the outcome of international environmental agreements. We aim to understand whether concerns for a critical level of damages induce cooperation among countries faced with the well-known free-riding problem, and yield sufficient emission reductions to avoid exceeding the threshold. Specifically, we focus on loss-averse countries negotiating under the threat of either high or low environmental damages. 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 than in the absence of reference dependence. We then introduce asymmetry, by allowing countries to differ in the degree of concern about the damages. We show that stable coalitions are mostly formed by the countries with higher threshold concerns. When enough countries exhibit standard preferences, the coalition size may diminish, regardless of the degree of concern by the others.

Keywords: Loss aversion; international environmental agreements; critical damages; threshold concerns

[1]School of Economics, Sogang University, GN622, Mapo-gu, Seoul, 04107, S. Korea.
Phone: +82-2-705-8505, Email: dorukiris@gmail.com
[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 Taehyun Ahn, Frank Venmans and participants at ASSET 2017, SIEP 2017, EAERE 2017, Osnabrück University, SED 2015, and Sogang University.

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 only 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 (DellaVigna, 2009; Barberis, 2013). Remarkably, loss aversion has not been used, to the best of our knowledge, in modeling environmental agreements.[1] Given the pervasiveness of reference point considerations in human decision-making, we investigate its role in affecting the

[1] One exception is İriş (2016). 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 and Vogt, 2003; Lange, 2006).

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 literature on dangerous climate change has recently focused on 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.

Inspired by the above, the premise of this work is that there exists a critical level of damages, which is viewed as a reference point separating a "safe" from a "dangerous" domain (Tavoni and Levin, 2014). Nations expect that business can carry on as usual below a given tolerable amount of environmental damage, according to the standard calculus of net benefits from pollution. However, above a critical level of damage from emissions, additional losses will ensue according to a multiplier effect.

In this paper we model threshold concerns by introducing a kink in the welfare function and allowing expectations about the damages to vary across countries.[2] Such asymmetry may arise for at least two reasons, both of which are compatible with our framework: (1) divergent views on the location of the threshold, possibly due to disagreement among the experts or the negotiators (Bosetti et al., 2017a,

[2] Karp and Simon (2013) also studies kinks on the welfare functions in IEAs.

Dannenberg et al., 2017); (2) differing evaluations of the losses incurred when exceeding the threshold. One can think of (1) as rooted in scientific uncertainty about its location, while the latter form of asymmetry can either be attributed to differences in the capital at stake (e.g. due to higher potential physical and financial losses from climate change in coastal areas), or in perceptions.

Here we concentrate on the latter behavioral interpretation, but remark that the findings detailed in Section 3 do not hinge on it and apply equally well to the other (non-behavioral) forms of asymmetry. Specifically, we study 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 critical level of admissible damages beyond which severe 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 below.

For tractability reasons, in Section 2 we abstract from the complexities arising from asymmetries in exposure to the damages from high concentrations of pollutants, and assume that countries are symmetric and agree on one value of the threshold, henceforth T. Introducing uncertainty on its location can destabilize cooperation, by removing the coordination equilibrium where (just) enough mitigation is undertaken to avoid steep losses. Under sufficiently large uncertainty, the game reverts to a prisoner's dilemma whose unique equilibrium is for all countries to eschew mitigation efforts (Barrett, 2013).

The related experimental literature on the provision of discrete public goods subject to provision thresholds corroborates this result.[3] 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

[3] See İriş and Tavoni (2016) for a recent review of the literature on tipping points and reference-dependent preferences in climate change games.

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 empirical and theoretical studies confirm the importance of leadership (Bosetti et al., 2017b; İriş et al., 2015; Marchiori et al., 2017), especially on the part of wealthy actors (Vasconcelos et al., 2014).

The model presented in Section 2 is an initial step in introducing realistic features in the standard coalition formation model of international environmental agreements. The symmetry and common knowledge assumptions utilized in Section 2 are likely to bias upwards the 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 introduce heterogeneous beliefs by letting a fraction of the countries perceive the critical level of damages to be higher than for the remaining countries.[4] This asymmetry in countries' threshold concerns may also exist due to heterogeneous beliefs on the location of the threshold.

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 critical damages, whose consequences would be felt equally by all countries.

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

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

degree of concern of the other countries. 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 symmetry. We begin by studying two benchmark cases, the games associated with non-cooperative behavior and full cooperation. We then 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. Section 4 discusses the implications of the main findings.

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, $e_i \geq 0$.[5] 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 objective benefit and damage functions:

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

[5] 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.

where β and γ are positive.

In addition to the standard calculus outlined above, in this section we assume that countries are identical also with respect to concerns about environmental damages. That is, each country i shares the same views on the magnitude of the critical threshold $T \geq 0$ representing the critical level of damages. Note that these concerns are based on country representatives' perceptions (which arguably reflect the views and biases of their domestic constituency), and thus differ from the objective damages in (1). If the level of environmental damages remains below the threshold, i.e., $D_i(E) \leq T$, then each country i enjoys being within the critical level of damages. If the level of environmental damages exceeds the threshold, $D_i(E) > T$, then each country's welfare drops due to the threat of sizeable damages. 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 (through increased emissions).

Put differently, governments anticipate a certain outcome (in terms of damages from climate change), and any deviation from this expected outcome is evaluated relative to the reference point. From the perspective of a loss-averse country, the marginal net benefit of emissions will be lower when the reference point is exceeded, and higher when damages are below the reference point.

The environmental gain-loss function of country i is written as follows:

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

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

Due to the global externality, the social welfare of a loss-averse country i depends on its own emissions, on the emissions of others, $e_{-i} = \{e_1, \dots, e_{i-1}, e_{i+1}, \dots, e_n\}$, and also on T. The social welfare of a loss-averse country i, $B_i(e_i) - D_i(E) + GL_i(E,T)$, can be expressed in the following general form:

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

$$w_i(e_i, e_{-i}, T) = \beta e_i - \frac{1}{2}e_i^2 - \frac{\gamma}{2}LE^2 + (L-1)T \qquad (3)$$

where L captures threshold concerns. It takes different values in the following three possible cases:

$$L(E) = \begin{cases} 1, \ no\ threshold\ \ (neutral\ domain) \\ 1+\alpha, \ D(E) \leq T\ \ (gain\ domain) \\ 1+\alpha\lambda, \ D(E) > T\ (loss\ domain) \end{cases} . \qquad (4)$$

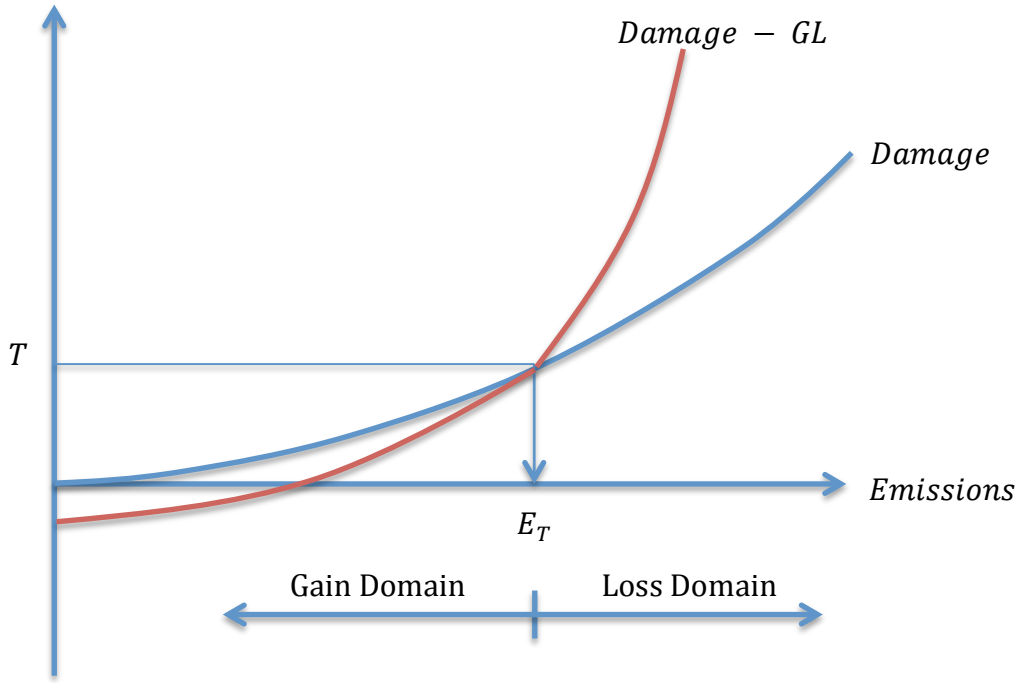
α is a positive scaling factor, measuring the degree to which country i cares about the environmental gain-loss function. If the level of environmental damages exceeds T, substituting $L = (1+\alpha\lambda)$ in (3) results in the loss domain. The gain domain results instead when $\lambda = 1$, implying $L = (1+\alpha)$. Similarly, the neutral domain is recovered by equating $\alpha = 0$, which implies $L = 1$.[7]

Note that the 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.

A loss-averse country's perceived damage incorporates the gain-loss function into the damage function and, thus, into 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.

[7] For some parameter values, country i's equilibrium emissions in the loss domain $(L = 1+\alpha\lambda)$ may yield damages below the threshold, thus suggesting that country i is in the gain domain. On the other hand, country i's equilibrium emissions in the gain domain $(L = 1+\alpha)$ may yield damages above the threshold. We can overcome this problem by extending the loss domain, assuming that countries maximize their welfare when they are in the loss domain in the ε-neighborhood of the critical threshold T. We discuss this matter in greater detail in Appendix 3.

Figure 1: Environmental Damage and Perceived Damage



Note: The loss-averse countries' perceived damage function (in red) is given by the difference between the damage and gain-loss functions (the damage function enters negatively and the gain-loss function enters positively into the social welfare). The loss-averse countries' perceived damage function is steeper in all domains and much steeper in the loss domain, compared to the standard damage function without threshold concerns (equation 1). This is due to the kink at T caused by the loss-aversion parameter λ.

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

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 the full cooperation case, all countries choose how much to jointly emit to maximize their aggregate social welfare function, $w = \sum_{i=1}^{n} w_i$. Remark 1 summarizes our findings of these cases.[8]

Remark 1 (Symmetric benchmarks):

- Non-cooperation:

[8] To increase readability, we avoid a significant amount of simple but tedious calculations in the paper. A Mathematica supporting file for these calculations is available at goo.gl/zypym3. Proofs are relegated to Appendix 1.

i. Country i's best-response function: $e_i\left(\sum_{j\neq i} e_j\right) = \frac{\beta-\gamma L(\sum_{j\neq i} e_j)}{1+\gamma L}$

ii. Per-country non-cooperative emission level: $e_{nc} = \frac{\beta}{1+n\gamma L}$

iii. Welfare under no cooperation: $w_{nc} = \frac{\beta^2(1-n\gamma L(n-2))}{2(1+n\gamma L)^2} + (L-1)T$

- Cooperation:

iv. Per-country cooperative emission level: $e_c = \frac{\beta}{1+n^2\gamma L}$

v. Welfare under full cooperation: $w_c = \frac{\beta^2}{2(1+n^2\gamma L)} + (L-1)T$

While both non-cooperative and cooperative emission levels, e_{nc} and e_c, 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 of w_{nc} and w_c 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.[9]

2.2. Partial Cooperation

The coalition formation game consists of three stages that are solved under the assumption that countries can look forward and infer backwards. Stage 1 is a participation game in which each country chooses 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

[9] 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.

member emitting e_s such that $E_s = se_s$. Each non-signatory emits e_{ns}, so that non-signatories collectively emit $E_{ns} = (n-s)e_{ns}$.

Non-signatories are Stackelberg followers: their behavior is described by the same best-response function as in the non-cooperative model (5). Signatories are the Stackelberg leaders: 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 non-signatories' best response function $e_{ns}(e_s)$. Remark 2 summarizes our findings for the case of partial cooperation.

Remark 2 (Partial cooperation):

- Non-signatories:
  - i. Best-response: $e_{ns}(e_s) = \frac{\beta - \gamma L s e_s}{X}$, where $X = 1 + \gamma L(n-s)$
  - ii. Emission level: $e_{ns} = \beta\left(1 - \frac{\gamma L n X}{\Psi}\right) = e_s + \frac{\beta \gamma L n (s-X)}{\Psi}$, where $\Psi = \gamma s^2 L + X^2$
  - iii. Indirect social welfare function:

$$\omega_{ns} = \beta^2\left(\frac{1}{2} - \frac{\gamma L(1+\gamma L)n^2 X^2}{2\Psi^2}\right) + (L-1)T$$

- Signatories:
  - iii. Emission level: $e_s = \beta\left(1 - \frac{\gamma L n s}{\Psi}\right)$
  - iv. Indirect social welfare function:

$$\omega_s = \beta^2\left(\frac{1}{2} - \frac{\gamma L n^2}{2\Psi}\right) + (L-1)T$$

- Aggregate:
  - v. Emission level: $E = E_s + E_{ns} = se_s + (n-s)e_{ns} = \frac{\beta n X}{\Psi}$

The non-signatory country i's best response depends on the joint emission of others countries, $(n-s-1)e_{ns} + se_s$, where, by symmetry, each non-signatory and also each signatory country emit the same level in equilibrium, e_{ns} and e_s respectively.

Note that we must restrict the parameters to guarantee that signatory and non-signatory emissions are positive, as there is no stock of emissions in the model. We

further restrict the parameters for emissions level to be higher for non-signatory countries than for signatories.

Proposition 1 (Conditions on emission levels):

i. $e_{ns} > e_s \Leftarrow s > X \Leftrightarrow \gamma < \frac{s-1}{(n-s)L}$

ii. $e_s > 0 \Leftrightarrow \gamma < \frac{4}{n\,L\,(n-4)}$ for $n > 4$

iii. $e_{ns} > 0 \Leftarrow \gamma < \frac{4}{n\,L\,(n-4)}$ 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 model of loss neutrality, with $L > 1$) additionally requires the relative impact of damages to be smaller. This condition plays an essential role on restricting the size of the stable coalition to be 2, 3, or 4 (see more on DS, pp.253).

The following Lemma, adapted from Proposition 2 in DS, defines the properties of indirect welfare functions.

Lemma 2 (Properties of indirect welfare functions—Proposition 2 in DS): Consider the indirect welfare functions of signatory and non-signatory countries, ω_s and ω_{ns}, respectively, and let $z^{min} = \frac{1+\gamma Ln}{1+\gamma L}$. 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 $s > z^{min}$ and $\omega_{ns}(s) < \omega_s(s)$ 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 2 shows that a country is better off as a signatory by more countries joining the coalition if the size of the coalition is small, and worse off as a signatory if the size of the coalition is large. Next, we discuss the impact of governments' threshold concerns on the welfare functions.

Proposition 3 (Accession under different threshold concerns): Let L' and L'' represent two different threshold concerns where $L'' > L'$, 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 3 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:

$$\omega_s(s^*) \geq \omega_{ns}(s^* - 1) \text{ and } \quad \omega_s(s^* + 1) \leq \omega_{ns}(s^*). \tag{6}$$

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,

[10] The conditions (6) 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).

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.

Proposition 4 (Effect of L on the stable coalition size): For $n > 4$, $\partial s^*/\partial L \geq 0$.

Proposition 4 is the main finding of the symmetric model, namely that the stable coalition size weakly increases with threshold concern. In Appendix 2 we illustrate this finding with a numerical example in which the size of the stable coalition increases from 2 to 3.

3. Asymmetric Model

In this section, we introduce heterogeneity in perceptions. Namely, as depicted in Figure 2, out of n countries, h have a high degree of concern for exceeding the threshold T and $n-h$ have low (or no) threshold concerns, $L_h > L_l$. Alternatively, as depicted in Figure 3, all countries share the same level of threshold concerns L, but they have divergent views on the location of the threshold T: h countries believe the threshold is low and $n-h$ countries believe it is high, $T_l \leq T_h$. Therefore, h countries are in the loss domain, and $n-h$ countries are either in the gain (or neutral) domain. Both interpretations are compatible because the only role played by T (or T_l and T_h) in the model is to determine in which domain the countries are, such that it does not affect the emission levels and the stability conditions (8-9).

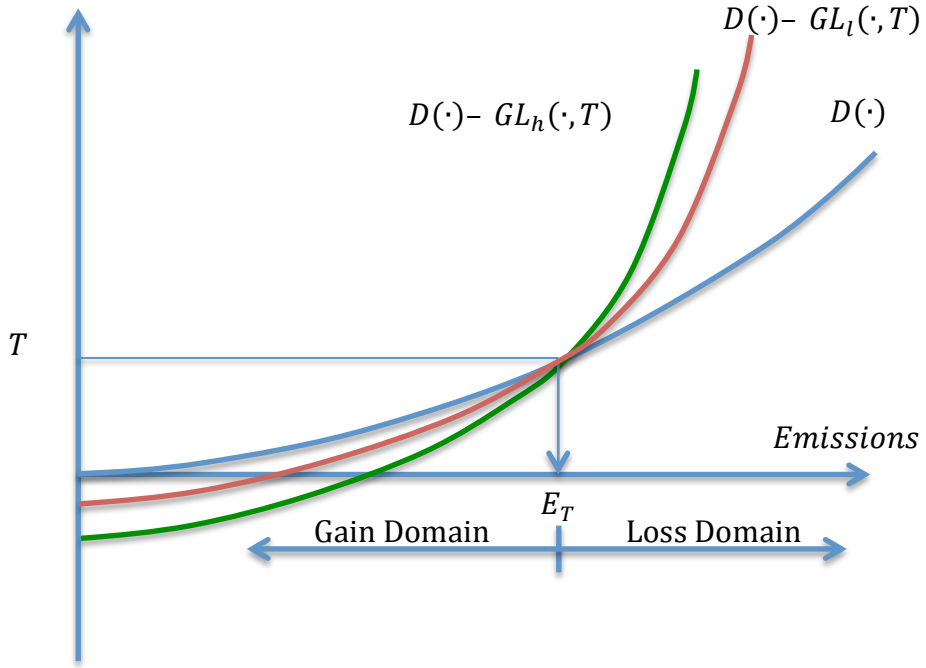
3.1. Two Benchmark Cases: The Non-cooperative and Full Cooperation Solutions

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:

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

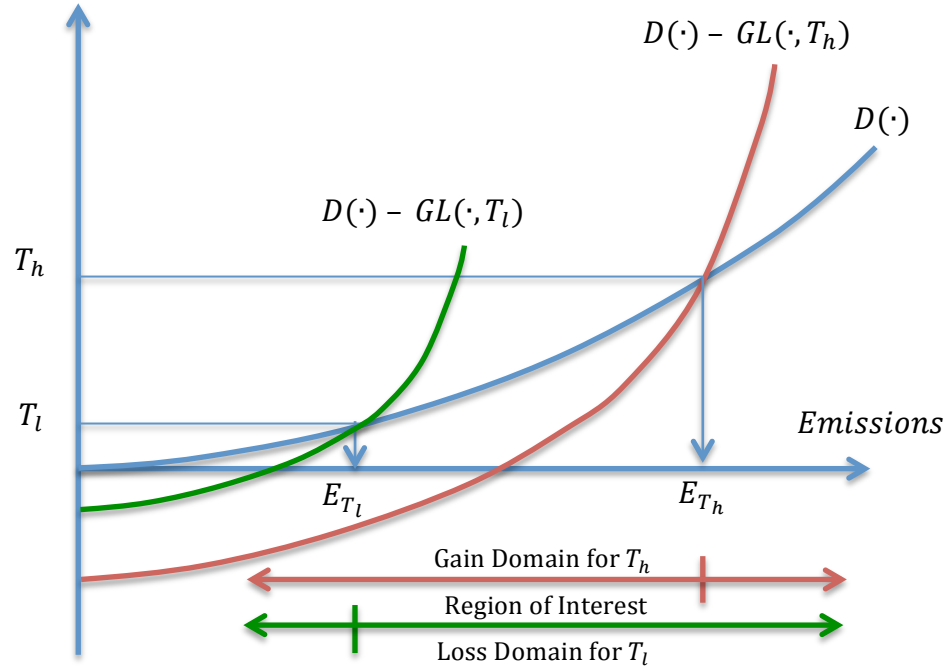
$$w_{li}(e_{li}, E, T) = \beta e_{li} - \tfrac{1}{2}e_{li}^2 - \tfrac{\gamma}{2} L_h(e_{li} + he_h + (n-h-1)e_l)^2 + (L_l - 1)T; \quad (7)$$

Figure 2: Asymmetric Threshold Concerns and Identical Beliefs on T



Note: Countries share the same beliefs on the location of the threshold, but their perceived damage functions are asymmetric: h countries have high threshold concerns (plotted in green) and n − k countries have low or no threshold concerns (plotted in red and blue, respectively).

Figure 3: Identical Threshold Concerns and Asymmetric Beliefs on T



Note: Countries have the same level of threshold concerns but they have asymmetric beliefs on T: h countries believe the threshold is low and n − h countries believe the threshold is high ($T_l \leq T_h$). Thus, h countries are in the loss domain (plotted in green), and n − h countries are either in the gain or neutral domain (plotted in red and blue, respectively).

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. In the full cooperation case, both types of countries jointly decide their emission levels to maximize their aggregate social welfare function, $w = \sum_{i=1}^{n} w_i$. Remark 3 summarizes the results for these polar cases:

Remark 3 (Benchmarks under asymmetry):

- Non-cooperation:
  - i. Country i's best-responses with high and low threshold concerns:

$$e_{hi}\big((h-1)e_h + (n-h)e_l\big) = \frac{\beta - \gamma L_h\big((h-1)e_h + (n-h)e_l\big)}{1+\gamma L_h}$$

$$e_{li}(he_h + (n-h-1)e_l) = \frac{\beta - \gamma L_l(he_h + (n-h-1)e_l)}{1+\gamma L_l}$$

  - ii. Best-responses under $e_{hi} = e_h$ and $e_{li} = e_l$ for any country i, with high and low threshold concerns:

$$e_h\big((n-h)e_l\big) = \frac{\beta - \gamma L_h(n-h)e_l}{1+\gamma h L_h}$$

$$e_l(he_h) = \frac{\beta - \gamma L_l h e_h}{1+\gamma L_l(n-h)}$$

  - iii. Equilibrium emission levels with high and low threshold concerns:

$$e_h^{nc} = \beta\left(\frac{1-\gamma(n-h)(L_h-L_l)}{1+\gamma(hL_h+(n-h)L_l)}\right) < \beta\left(\frac{1+\gamma h(L_h-L_l)}{1+\gamma(hL_h+(n-h)L_l)}\right) = e_l^{nc}$$

  - iv. $e_h^{nc} > 0 \Leftarrow \gamma < \frac{1}{(L_h-L_l)(n-h)}$

- Cooperation:
  - v. Equilibrium emission levels with high and low threshold concerns:

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

  - vi. $e_h^c = e_l^c < e_h^{nc} \Leftrightarrow \gamma < \frac{(h-1)L_h+(n-h)L_l}{(hL_h+(n-h)L_l)(L_h-L_l)(n-h)}$

Notice that the denominators of both emission levels in (iii) 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: $e_h^{nc} < e_l^{nc}$.

Under full cooperation, both types of countries emit the same emission levels. (iv) and (vi) are similar to Proposition 1 (iii). They additionally require the difference between high and low threshold concerns to be limited for $e_h^{nc} > 0$ and for $e_h^{nc} > e_h^c = e_l^c$.

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 $n - h - s_l$. We denote the emission levels of signatories and non-signatories with high (L_h) and low (L_l) threshold concerns as e_{sh}, e_{sl}, e_{nsh} and e_{nsl}, respectively. Proposition 6 summarizes the results for the case of partial cooperation:

Remark 4 (Emissions and welfare under partial cooperation among asymmetric countries):

- Best responses of non-signatories with high and low threshold concerns:
  - i. Non-signatory country i with high threshold concerns:

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

  - ii. $e_{nshi} = e_{nsh}$ and $e_{nsli} = e_{nsl}$ for any country i:

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

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

iii. Since non-signatories decide simultaneously:

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

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

- Emission Levels:

iv. Signatories with high and low threshold concerns:

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

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

v. Since signatories decide simultaneously:

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

vi. Non-signatory countries with high and low threshold concerns:

$$e_{nsh} = \frac{\beta(\Omega - \gamma L_h nY)}{\Omega} \leq \frac{\beta(\Omega - \gamma L_l nY)}{\Omega} = e_{nsl}$$

vii. $e_{sh} = e_{sl} < e_{nsh} \Leftarrow \gamma < \frac{(s_h - 1)L_h + s_l L_l}{(h - s_h)L_h + (n - h - s_l)L_h L_l}$

$$e_{sh} = e_{sl} > 0 \Leftarrow \gamma < \frac{1}{(s_h L_h + s_l L_l)(n - h - s_l)}$$

viii. Aggregate emissions, $E^A = E_{sh} + E_{sl} + E_{nsh} + E_{nsl}$:

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

- Indirect social welfare functions of signatories and non-signatories with high and low threshold concerns:

$$\omega_{sh} = \beta^2\left(\frac{1}{2} - \frac{\gamma n^2\big(Y^2 L_h + \gamma(s_h L_h + s_l L_l)\big)}{2\Omega^2}\right) + (L_h - 1)T$$

$$\omega_{sl} = \beta^2\left(\frac{1}{2} - \frac{\gamma n^2\big(Y^2 L_l + \gamma(s_h L_h + s_l L_l)\big)}{2\Omega^2}\right) + (L_l - 1)T$$

$$\omega_{nsh} = \beta^2\left(\frac{1}{2} - \frac{\gamma n^2 Y^2 L_h(1 + \gamma L_h)}{2\Omega^2}\right) + (L_h - 1)T$$

$$\omega_{nsl} = \beta^2\left(\frac{1}{2} - \frac{\gamma n^2 Y^2 L_l(1 + \gamma L_l)}{2\Omega^2}\right) + (L_l - 1)T$$

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:

$$\omega_{sh}(s_h^*, s_l^*, h, n) \geq \omega_{nsh}(s_h^* - 1, s_l^*, h, n), \omega_{sh}(s_h^* + 1, s_l^*, h, n) \leq \omega_{nsh}(s_h^*, s_l^*, h, n); \quad (8)$$

$$\omega_{sl}(s_h^*, s_l^*, h, n) \geq \omega_{nsl}(s_h^*, s_l^* - 1, h, n), \omega_{sl}(s_h^*, s_l^* + 1, h, n) \leq \omega_{nsl}(s_h^*, s_l^*, h, n). \quad (9)$$

Due to the asymmetry, these conditions depend on the number of signatories with high and low threshold concerns: s_h^* and s_l^*, and their sum yield the stable size of coalition. 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 concerns 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 (9) 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 (8) 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			
s_l/s_h	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	hl			l	hl		l	l	hl	l	l	l	hl
1						h			h			l	h				h				h	
2			hl		hl			hl				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, while some countries with high threshold concerns stay outside of the coalition. It is difficult to disentangle effects and understand the reasons behind this surprising result, which deserves future investigation.

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			
s_l/s_h	0	0	1	0	1	2	3	0	1	2	3
0			l		l	hl			l	l	hl
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**			
s_l/s_h	0	0	1	0	1	2	3	0	1	2	3	0	1	2	3
0						hl	l			hl	l			l	hl
1			l		l	h			l	h			l	h	
2	l	l	h	l	h				h				h		
3		h		h				hl				hl			

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

In ecological processes, threshold uncertainty is often irreducible; nevertheless, scientists 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 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.

While we do away with the complexities arising from explicitly modeling how uncertainty muddles the value of the threshold, our framework allows for divergent views on its location, which may arise due to scientific or political disagreement among the experts or the negotiators (Bosetti et al., 2017a, Dannenberg et al., 2017). Under this interpretation, countries disagree because of scientific uncertainty about the location of the threshold, or because of the difficulty in translating a given threshold into the effort required to avoid overstepping such a boundary, as argued in Barrett (2013). Alternatively, countries may have different evaluations of the losses incurred when in the "danger zone". The asymmetric behavior modelled here may thus be attributed to differences in the capital at stake (e.g. due to higher potential physical and financial losses from abrupt climate change in coastal areas), or in perceptions.

In this modified IEA setup, we ask whether the traditionally negative prediction of either small or ineffective international environmental agreements can be reverted (Barrett, 1994; Carraro and Siniscalco, 1993). Specifically, we study 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) threshold level of damages from emissions of pollutants linked with industrial production, beyond which more 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 and the threat it represents. 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.

We show that, both under full cooperation and when all countries act non-cooperatively, threshold concerns reduce global emission levels relative to the standard model, even though it does not necessarily increase countries' welfare, either. We further establish 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 countries with threshold concerns are more likely to take significant environmental decisions on reducing their emissions, provided that their governments and delegates 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.[11] Calibrating the model with parameters

[11] Recent trends of increasing mobility of capital and labor and global economic growth driven by emerging markets, which are abstracted as benefits of emission, could change quantitative results but the qualitative results presented here would remain the same.

estimated from real-world data and empirical evidence appears to be a fruitful avenue for testing the stylized model we have introduced here.[12] 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 increasingly large damages from unabated emissions 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.

[12] For instance, one can estimate country-specific costs and benefits of emissions, as well as experimental findings on loss aversion in representative countries.

Bosetti, V., Weber, E., Berger, L., Budescu, D., Liu, N. and Tavoni, M., 2017. COP21 climate negotiators' responses to climate model forecasts. *Nature Climate Change*, 7(3).

Bosetti, V., Heugues, M. and Tavoni, A., 2017. Luring others into climate action: coalition formation games with threshold and spillover effects. *Oxford Economic Papers*, 69(2), 410-431.

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

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.

Dannenberg, A., Zitzelsberger, S. and Tavoni, A., 2017. Climate negotiators' and scientists' assessments of the climate negotiations. Nature Climate Change, online first.

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.

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

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.

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. 2016. Economic targets and loss-aversion in international environmental cooperation. Journal of Economic Surveys 30 (3), 624-648.

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.

Karp, L. and Simon, L. (2013). Participation games and international environmental agreements: A non-parametric model. Journal of Environmental Economics and Management 65 (2), 326-344.

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.

Lange, A., Vogt, C., 2003. Cooperation in international environmental negotiations due to a preference for equity. Journal of Public Economics, 87 (9-10), 2049-2067

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

Marchiori C, Dietz S, Tavoni A (2017) Domestic politics and the formation of international environmental agreements. Journal of Environmental Economics and Management, 81, 115–131

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.

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), 2212-2216.

Appendix 1: Proofs

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 Proposition 1:

i. Straightforward from Remark 2, ii.

ii. From Remark 2, 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 $\underline{s} = argmin_s\, A(s) = \frac{2+n+2\gamma Ln}{2(1+\gamma L)}$. For $A(s) > 0$ for any s, it is sufficient to show that $A(\underline{s}) > 0$. One can easily find that $A(\underline{s}) = \frac{4-\gamma L(n-4)n}{4(1+\gamma L)}$ and for $A(\underline{s}) > 0$, we need $\gamma < \frac{4}{n\,L\,(n-4)}$.

iii. From Remark 2, 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 $\Phi(\overline{s}) = \left(\frac{2+\gamma Ln(2+\gamma L)}{2(1+\gamma L)}\right)\left(\frac{2-\gamma^2 L^2 n}{2(1+\gamma L)}\right) + \left(\frac{\gamma L(2+\gamma Ln)^2}{4(1+\gamma L)^2}\right)$ 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 $\Phi(\overline{s}) = \frac{4+20\gamma L-25\gamma^3 L^3}{4(1+\gamma L)}$ and for $\Phi(\overline{s}) > 0$, the following condition should hold: $25\gamma^3 L^3 - 20\gamma L - 4 < 0$ and indeed holds for $\gamma < \frac{4}{5L}$.

Proof of Lemma 2:

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 3:

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 2, 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 2 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 2 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 4: 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.

Appendix 2: Numerical Example

Let us 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 4 depicts the case when governments do not exhibit concerns for dangerous climate change beyond a tipping point, $L = 1$. Figure 5 focuses instead on countries with some degree of threshold concern: we set $L = 1.5$ for visual clarity.[13] While T does not play any role in Figure 4, T is set to be 1 in Figure 5, which places countries in the loss domain.[14] 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, $s = 1, \dots, 4$.

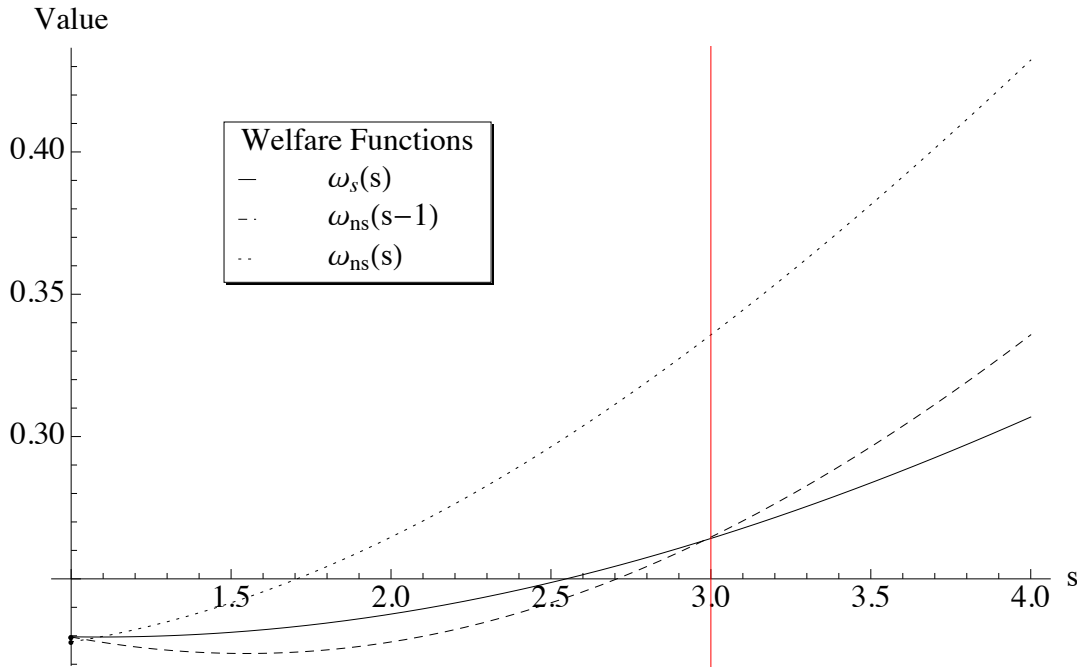
In Figure 4, one can observe that coalition size $s^* = 2$ is internally stable, $\omega_s(s^*) \geq \omega_{ns}(s^* - 1)$, since the solid curve is above the dashed curve at $s = 2$. Note also that these two curves intersect at $s = 2.976$, so $s = 3$ is not internally stable. Moreover, coalition size $s^* = 2$ 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 5, 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.

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

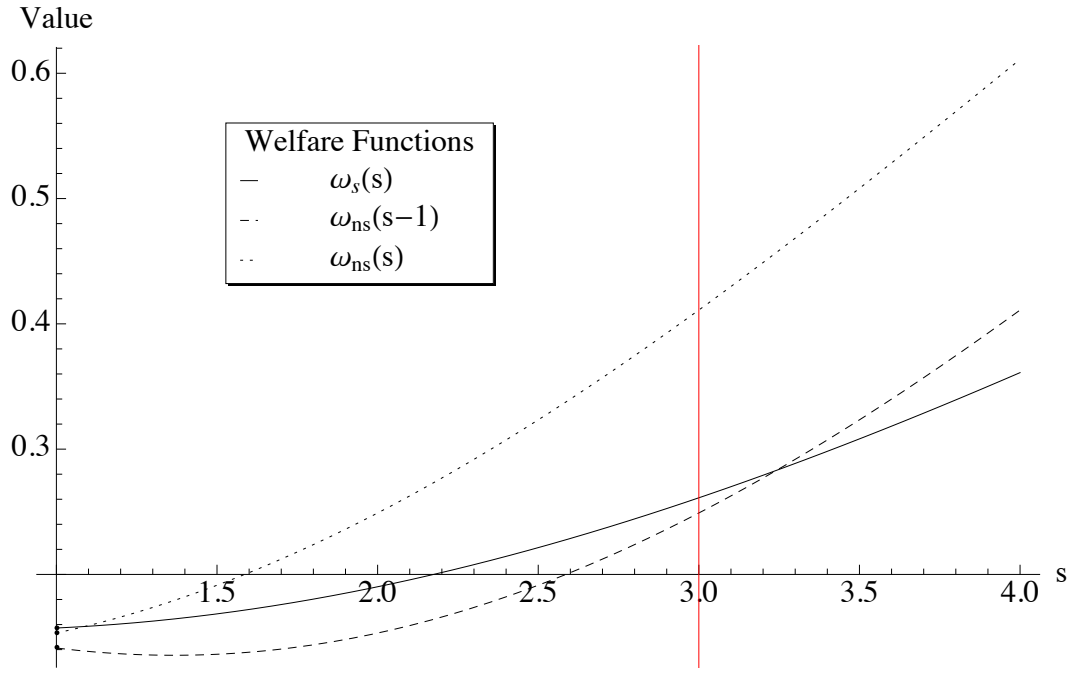
[14] Remember that T does not affect the emission levels, once the domain is determined. It does levy the welfare level, but in equal measure for all indirect welfare functions $\omega_s(s)$, $\omega_{ns}(s)$, and, $\omega_{ns}(s-1)$. Thus, the size of the coalition does not depend on T so long as countries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

Figure 4: Coalition Size without Threshold Concerns ($L = 1$)
$n = 10, \beta = 5/3, \gamma = 0.01$.



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

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



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

Appendix 3: A Technical Note

Here we discuss the technical issues noted in footnote 6. Let us first give an example to clarify the problem. Suppose the critical threshold is $T = D\left(\frac{n\beta}{1+n\gamma(1+\alpha)}\right) - \varepsilon$, where $\varepsilon > 0$, then

i. In the case the equilibrium lies in the gain domain we have the aggregate emission level $\frac{n\beta}{1+n\gamma(1+\alpha)}$, leading damages higher than T

ii. In the case the equilibrium lies in the loss domain we have the aggregate emission level $\frac{n\beta}{1+n\gamma(1+\alpha\lambda)}$, leading damages lower than T for ε sufficiently low.

As noted in footnote 6, we can eliminate this problem by extending the loss domain. Specifically, by assuming that countries maximize their welfare as they are in loss domain in the ε-neighborhood of the critical threshold T. Incorporating this modifies environmental gain-loss function (2) as follows:

$$GL_i(E,T) = \begin{cases} T - D_i(E), & D_i(E) \le T - \varepsilon \\ \lambda\big(T - D_i(E)\big), & D_i(E) > T - \varepsilon \end{cases} \qquad (2')$$

Similarly, this will modify equations (3-5). Nevertheless, T and ε disappear once the first order condition taken and, thus, this modification does not affect the solution of the problem.