# Shaping Climate Cooperation: The Role of Global Environmental Norms

Effrosyni Diamantoudi, Concordia University
Olha Hnatyshyn, Concordia University
Eftichios S Sartzetakis, University of Macedonia*

June 5, 2025

**Abstract**

Game-theoretic analyses of International Environmental Agreements (IEAs) find that self-interested countries fail to form large, stable coalitions capable of significantly reducing global emissions due to strong free-riding incentives. This paper extends the analysis by incorporating normative/moral factors into countries welfare functions. We assume that governments demonstrating genuine environmental commitment gain domestic support from environmentally conscious citizens and enhance their international reputation. A country that exceeds a global environmental norm experiences a welfare gain (e.g., pride, "warm glow"), while falling below incurs welfare losses (e.g., shame, reputational cost). In a simultaneous-move game, incorporating these normative elements does not increase the size of stable coalitions but leads both signatory and non-signatory countries to reduce emissions and enjoy higher welfare. When the coalition acts as a Stackelberg leader, both coalition size and emissions reductions improve. Aggregate emissions decline more rapidly as countries' responsiveness to global norms increases, especially when non-signatories are highly sensitive to normative pressures. Our results suggest that normative motivations can significantly enhance environmental outcomes even without enlarging coalitions. Strengthening global norms and increasing countries' sensitivity to them—particularly among non-signatories—can promote broader cooperation and more effective climate action.

*Corresponding author: E.S. Sartzetakis, Department of Economics, University of Macedonia, 156 Egnatia Str., 54636. Email: esartz@uom.edu.gr

# 1 Introduction

In the contemporary global landscape, the intertwining of economic and ethical incentives has become a pivotal force shaping national environmental policies and practices, and fostering international cooperation. Particularly in the realms of environmental sustainability and public accountability, the economic motivations of countries are increasingly shaped by ethical considerations and public pressure. The urgency of environmental issues, such as climate change, biodiversity loss, and pollution, has catalyzed a paradigm shift in how countries formulate and implement economic policies. Traditionally, economic growth was pursued with little regard for environmental degradation, often resulting in significant ecological harm. However, the rising public awareness and advocacy for sustainable development have pressured governments to integrate ethical considerations into their economic frameworks, adopt more stringent environmental standards and practices and promote international cooperation aimed at mitigating environmental harm.

In recent decades, the urgency of global environmental challenges, particularly climate change, has spurred the creation of numerous international environmental agreements. While some, like the Montreal Protocol, have seen great success, others have failed to effectively address the issues they were designed to resolve. Climate change remains the most prominent example, where free-riding incentives continue to undermine the global cooperation needed for meaningful progress. The success of international environmental agreements often hinges on the delicate balance of competing priorities among participating countries. Aligning national interests with global objectives is crucial for fostering greater cooperation and encouraging countries to collectively tackle pressing environmental issues.

The role of public pressure in this context cannot be underestimated. In recent years, citizens — empowered by unprecedented access to information and communication tools — have become more vocal in demanding that their governments take action on global environmental issues. This growing demand for accountability has already led some countries, or a group of countries like the European Union, to implement more stringent environmental policies, thereby enhancing their international reputation and positioning themselves as leaders in the global push for sustainability. By aligning economic objectives with ethical imperatives at both the national and international levels, countries can not only gain the trust and support of their citizens but also attract environmentally responsible investments and elevate their global standing. This paper explores the case in which countries, responding to public pressure, incorporate an ethical dimension into their decision-making process. In addition to considering material benefits and damages from emissions, countries factor in the evaluation of their environmental standing relative to other nations. We then analyze how these ethical considerations influence emission choices and the level of cooperation among countries.

There is extensive literature that incorporates ethical dimensions of climate change, moral behavior and altruism of countries into economic models to study the stability of international environmental agreements. In their paper Dernbach and Brown (2008) emphasize that nations must consider the interests of others and limit their emissions based on principles of fairness to achieve a global solution. They argue that framing the issue ethically has multiple benefits: it shifts the motivation for climate action from self-interest to a broader responsibility towards others, and it highlights the necessity of active citizen engagement for the success of climate initiatives. Daube and Ulph (2016) introduce a novel perspective on moral behavior, suggesting that countries recognize the disadvantages of not acting in their self-interest. The authors adhere to Kantian ethics, in which the morality of an action is evaluated based on conformity to established principles or duties. They argue that genuinely moral actions arise from a sense of duty and adherence to moral law rather than being motivated by personal inclination or pursuit of specific outcomes. The authors illustrate that embracing a Kantian approach encourages significant altruism among individuals, resulting in substantial reductions in the consumption of environmentally harmful goods. Rashid (2023) examines the ethical challenges linked to greenhouse gas emissions, emphasizing deficiencies in current theoretical frameworks and underscoring the pressing necessity for global cooperation.

Gróning, and Peters (2010) demonstrate that when countries prioritize justice and fairness in their preferences, the tendency to free-ride on signatories diminishes. This enhances the effectiveness of participating in international environmental agreements, leading to the larger size of stable coalition. Eichner and Pethig (2024a) investigate the impact of moral behavior on coalition formation. The authors introduce homo moralis preferences enabling countries to base their decisions on the material benefits assuming that all other countries act similarly with a predetermined probability. The authors examined two distinct dimensions of moral considerations, moral behavior regarding the membership in the agreement and moral conduct regarding emissions reduction. The former involves the notion that countries might feel a moral duty to join and participate in the environmental agreement.

When countries perceive participation as a moral imperative, they may be inclined to form larger coalitions than they would if driven solely by self-interest. Moral behavior regarding emissions reduction involves countries taking into account the moral ramifications of their emissions levels. In this context, countries may choose to curtail their emissions because they regard it as a moral obligation, regardless of what other countries are doing. The authors conclude that while moral considerations regarding membership tend to enlarge coalitions, emissions morality tends to reduce coalition size. However, the interplay between these opposing influences does not straightforwardly determine the stable size of a coalition. The final coalition size depends on the relative strength of these moral considerations and how they interact with one another.

In their most recent work, Eichner and Pethig (2024b) examine games involving countries that exhibit Kantian moral behavior within the International Environmental Agreements (IEA) context. Authors distinguish between emissions and membership moralists, where moralists' welfare blends characteristics of Kantians and materialists. In games with emissions moralists, stable coalitions are no larger than those with materialists. In contrast, games with membership moralists show that coalition stability increases with higher membership morality. Overall, higher moral behavior among either type of moralists leads to a reduction in the aggregate emissions.

The role of altruism in enhancing cooperation is explored by Van der Pol et al. (2012) and most recently by Schopf (2023) and Sacco (2024). Sacco (2024) introduces altruistic stability conditions that enable the formation of large coalitions. Through simulations, the author demonstrates that the most effective pollution abatement outcomes are achievable with coalitions that include a majority of countries, rather than necessitating universal participation. Van der Pol et al. (2012) explores two types of altruism in countries' decision to participate in the agreement, impartial altruism, where concern extends to all countries, and community altruism, limited to coalition partners. The authors conclude that even a modest degree of altruism can stabilize the Grand Coalition. Schopf (2023) extends model of Van der Pol et al. (2012) by incorporating altruism into both stages of the game. The author finds that the coalition size never exceeds six countries, and the aggregate emissions and global welfare closely align with non-cooperative values. Nevertheless, altruism narrows the gap between the individually and the socially optimal emissions, thereby increasing total gains.

Breton and Sbragia (2023) examine the stability of international environmental agreements aimed at achieving a shared emissions reduction target. Authors introduce a concept of self-image, i.e. countries place value on their self-image resulting from their commitment to these agreements, and demonstrate that when countries prioritize their self-image, stable agreements with meaningful participation can be established, which effectively reduce the aggregate pollution and enhance global welfare. Furthermore, authors suggest that a broad collective effort, driven by modest ambitions, yields better results than a scenario where only a few countries commit to a significant reduction of their emissions.

The present paper contributes to the existing literature by examining the case in which countries consider their global environmental standing, reflecting their citizens' environmental awareness. We assume that countries incur ethical or moral costs when their environmental behavior falls short of a global norm, while they gain benefits when their environmental achievements surpass this norm. Specifically, we extend the standard model, in which each country maximizes the benefits from its own emissions net of damages from global emissions, by introducing a function that reflects the evaluation of a country's deviation from the global emissions norm. This extension enables us to analyze how incorporating moral imperatives into countries' decision-making processes affects both emission choices and the level of environmental cooperation. We assume that the global environmental norm is the average level of global emissions. Following the standard approach, we model environmental cooperation as a two-stage game: in the first stage, countries decide whether to join a coalition whose members internalize the external damages of their emissions on all coalition members, and in the second stage, they choose their level of emissions. We consider both the case in which countries choose their emission levels simultaneously and the case in which coalition members act as a leader.

We find that when countries incorporate ethical/moral motivations into their decision-making, it generally leads to significant reductions in aggregate emissions and improvements in overall welfare. However, it does not always result in increased levels of cooperation. Specifically, in the simultaneous game, when coalition members and free-riders share the same level of ethical motivation, emissions decrease for both groups, yielding substantial welfare benefits. Yet, the introduction of ethical motivations does not improve cooperation levels. On the other hand, when coalition members are more ethically motivated than outsiders, the size of the coalition shrinks.

In the leadership game, when both groups are equally motivated, they choose lower emissions, and the coalition size increases compared to the scenario without ethical motivations. However, if coalition members' ethical motivations are stronger than those of the free-riders, the coalition size decreases as the difference in motivations grows. For large differences in ethical motivations, emissions may increase, and we could even see a reduction in aggregate welfare.

In summary, when all countries share the same level of ethical motivation, they opt for lower emissions, even though cooperation does not increase significantly in most cases. However, if the ethical motivations of non-member countries are lower than those of coalition members, the welfare benefits diminish. Particularly in the case where the coalition acts as a leader, improving the ethical motivations of coalition members is not highly effective if free-riders do not enhance their own ethical considerations. Therefore, it is much more effective to persuade non-member countries to become more aligned with the global emissions norm, emphasizing the potential benefits they could gain by doing so.

The remainder of the paper is organized as follows. Section 2 presents the model. Section 3 analyzes the equilibrium outcomes in the simultaneous coalition formation game. Section 4 examines the equilibrium solutions in the leadership coalition formation game. Section 5 illustrates the results using quadratic functions. Section 6 compares the results of the models and concludes the paper.

## 2 The Model

We assume $n$ symmetric countries, $N = \{1, 2, 3, ..., n\}$, whose production and consumption activities generate greenhouse gas emissions responsible for climate change. The accelerating impacts of climate change—most notably the rising frequency and severity of extreme weather events—are already visible. Coupled with widespread information campaigns, these developments have substantially heightened public awareness. As a result, governments are facing growing scrutiny regarding their climate policies and actions. Environmentally conscious citizens closely monitor their government's environmental commitments, often comparing them with those of other nations.

A government that demonstrates genuine dedication to environmental sustainability earns greater public trust and support, while also enhancing its international reputation. Consequently, environmental considerations are becoming an integral part of national decision-making, shaped by the need to respond to public expectations and align with evolving global standards.

The emergence of various environmental and climate performance indices[1] has made cross-country comparisons more accessible and transparent. Moreover, research has shown that some of these indices serve as reliable predictors of economic risk.[2] This suggests that governments are increasingly motivated to enhance their environmental performance not only to meet public demands, but also to attract investment by presenting themselves as stable, low-risk, and ethically responsible actors on the global stage.

According to the above, a country that is able to exceed a given global norm experiences a welfare increase (sense of pride, warm glow), while falling behind the norm decreases its welfare (sense of shame, loss of status). This assumption captures the sentiment of environmentally concerned citizens, who experience reduced utility when their country's emissions surpass the global standard, thereby influencing the nation's decision-making process. Let $\widehat{e}$ represent the global norm and $\psi_i$ denote the global norm function, i.e. the deviation of country $i$'s environmental performance from this norm, $\psi_i = e_i - \hat{e}$. The effect of deviating from the global norm on country $i$'s welfare–negative or positive depending on whether $\psi_i \gtreqless 0$–is denoted by $C_i\left(\psi_i\right)$. This effect motivates countries to respond to global efforts aimed at reducing global pollution levels.

As a result, country $i$'s social welfare is determined by the benefits derived from economic activities, $B_i\left(e_i\right)$, minus the sum of damages associated to global pollution, $D_i\left(E\right)$, and the cost of falling behind the global norm, $C_i\left(\psi_i\right)$. Since both benefits from own emissions and the global norm effect are behavioral elements, we assume that are concave functions of their arguments, $B' \geq 0$, $B^{''} < 0$ and $B\left(0\right) = 0$, $C' \geq 0$, $C^{''} < 0$ and $C\left(0\right) = 0$. We further assume that environmental damages are convex in global emissions, $D^{'} \geq 0$, $D^{''} > 0$ and $D\left(0\right) = 0$. Assuming completely symmetric countries, country $i$'s social welfare is,

$$W_i\left(e_i, \psi_i, E\right) = B\left(e_i\right) - D\left(E\right) - C\left(\psi_i\right), \tag{1}$$

where $E$ is the aggregate level of pollution, $E = \sum_{i=1}^{n} e_i$.

The notion of a global norm in emissions can be conceptualized in multiple ways.

[1] Among others, the Environmental Performance Index by Country (EPIC), the Environmental Performance Index (EPI), the Environmental Country Reviews (EPCR), the Sustainable Development Report (SDR), and the Climate Change Performance Index (CCPI).

[2] Peiró-Signes et al. (2022) identified a positive and statistically significant correlation between the Environmental Performance Index (EPI) and the Country Risk Score (CRS). This finding indicates that countries enhancing their sustainability and environmental performance are more likely to offer stable, low-risk, and ethically sound conditions for investment.

It may be represented by a fixed benchmark—such as pre-industrial emission levels—or a theoretical first-best outcome. Alternatively, it can take the form of a dynamic, endogenous standard that evolves with global emissions. For instance, Breton and Sbragia (2023) define the norm as a linear function decreasing in aggregate emissions, reflecting shifting expectations as overall emissions change.[3]

In this paper, we define the global emissions norm as the average per-country emissions level, $\hat{e} = \frac{E}{n}$. This formulation is analogous to status-based utility models, where an agent's relative position matters. Accordingly, each country's deviation from the norm is captured by the function $\psi_i = e_i - \frac{E}{n}$, where $e_i$ denotes emissions by country $i$, and $E$ is total global emissions.

The formation of an international environmental agreement is modeled as a two-stage, non-cooperative game. In the first stage, each country decides whether to join the agreement. In the second stage, all countries—signatories and non-signatories—choose their emissions levels. We assume that a single coalition forms, comprising a subset of countries $S \subset N$, while the remaining countries in$N \backslash S$ remain outside the agreement. Let $S = |s|$ denote the number of coalition members.

Coalition members are assumed to adopt a common emissions level, denoted by $e_s$ while each non-member emits $e_{ns}$, while non-members emit, $e_{ns}$.The resulting aggregate emissions are therefore given by, $E = E_s + E_{ns} = se_s + (n - s)\, e_{ns}$ . We assume that all countries care about their environmental status—or self-image—relative to the global emissions norm. Specifically, countries incur a welfare loss when their emissions exceed the global norm, $\hat{e} = \frac{E_s + E_{ns}}{n}$, and experience a welfare gain when their emissions fall below it.

In the second stage—the emissions game—coalition members choose a common emissions level by jointly maximizing their aggregate welfare. In contrast, non-signatories choose their emissions levels individually, optimizing their own welfare without internalizing the externalities their emissions impose on others. We analyze both the case in which signatories and non-signatories choose emissions simultaneously, and the case in which the coalition acts as a leader. The former case is examined first in the next Section.

In the first stage—the membership game—countries decide whether to join the coalition

[3] In particular, the self-image function, as the norm function is denoted in Breton and Sbragia (2023), is $S_{it}(x_{it}, P_t) = \frac{r_0 - r_1 P_t}{\alpha} - x_{it}$, where $x_{it}$ is country $i$'s emissions and $P_t$ aggregate emissions at time $t$. Since the self-image (norm) component enters positively in country's welfare it is similar to the present paper's specification. However, in Breton and Sbragia (2023) the coalition does not maximize members' aggregate welfare but rather agrees on a common emissions target.

or free-ride on its mitigation efforts. To determine the size of the stable coalition, we apply the internal and external stability criteria originally proposed by D'Aspremont et al. (1983), which have been widely used in the literature. A coalition is internally stable if no member has an incentive to leave, and externally stable if no non-member has an incentive to join. Denoting by $s^*$ the size of the stable coalition, the two stability conditions can be formally expressed as follows,

$$\hat{W}_s\left(s^*\right) \geq \hat{W}_{ns}\left(s^*-1\right) \ , \tag{2}$$

$$\hat{W}_s\left(s^*+1\right) \leq \hat{W}_{ns}\left(s^*\right) \ . \tag{3}$$

In what follows, we first employ general functional forms—consistent with the properties outlined above—to analyze the effects of global norm considerations, beginning with the simultaneous-move game and subsequently extending the analysis to the leadership setting. We then specify particular functional forms to numerically illustrate and validate the theoretical insights derived from the general framework.

# 3 Simultaneous coalition formation game

In this section, we consider a setting in which both signatories and non-signatories choose their emission levels simultaneously. The coalition acts collectively, selecting its optimal emissions level—denoted by $e_s^*\left(e_{ns}\right)$—to maximize the aggregate welfare of its members, taking as given the emission level of non-signatories, $e_{ns}$,

$$\max_{e_s} sW_s = \max_{e_s}\left\{s\left(B\left(e_s\right) - C\left(\psi_s\right) - D\left(E\right)\right)\right\},$$

where $\psi_s = e_s - \hat{e}$, $\hat{e} = \frac{E}{n}$ and $E = se_s + (n-s)e_{ns}$.

The first-order condition of the above maximization problem is,

$$\frac{\partial B\left(e_s\right)}{\partial e_s} = \frac{\partial C\left(\psi_s\right)}{\partial\left(\psi_s\right)}\frac{\partial\left(\psi_s\right)}{\partial e_s} + \frac{\partial D\left(E\right)}{\partial E}\frac{\partial E}{\partial e_s}. \tag{4}$$

Solving (4) yields signatories' level of emissions, $e_s^*\left(e_{ns}\right)$, as a function of non-signatories emissions.

In contrast, non-signatories choose their emission levels independently, aiming to maximize individual welfare without internalizing the externalities imposed on others,

$$\max_{e_{ns}} W_{ns} = \max_{e_{ns}} \left(B\left(e_{ns}\right) - C\left(\psi_{ns}\right) - D(E)\right),$$

where $\psi_{ns} = e_{ns} - \hat{e}$, $\hat{e} = \frac{E}{n}$ and $E = se_s - (n-s-1)e_i + e_{ns}$.

The solution of the first-order condition of the above maximization problem,

$$\frac{\partial B(e_{ns})}{\partial e_{ns}} = \frac{C(\psi_{ns})}{\partial\left(\psi_{ns}\right)} \frac{\partial\left(\psi_{ns}\right)}{\partial e_{ns}} + \frac{\partial D\left(E\right)}{\partial E} \frac{\partial E}{\partial e_{ns}}\ , \tag{5}$$

yields non-signatories emissions level, $e^*_{ns}\left(e_s\right)$, as a function of signatories emissions.

Two key results emerge directly from the first-order conditions. First, all countries reduce their emissions relative to a scenario without global norm considerations. This outcome arises because countries experience a loss in welfare when their emissions exceed the global benchmark, and thus incorporate this self-image cost in their decision-making. The degree of emissions reduction differs between signatories and non-signatories, even when both groups exhibit the same sensitivity to the global norm. This asymmetry results from the fact that coalition members internalize the impact of their emissions on the welfare of all other coalition members, while non-signatories act unilaterally and ignore such externalities.

To see this, recall that the marginal cost of deviating from the global norm is increasing, that is, $\frac{C(\psi_i)}{\partial(\psi_i)} > 0$ and $\frac{\partial \psi_i}{\partial e_i} = 1 - \frac{\partial \hat{e}}{\partial e_i} > 0$, $i = s, ns$, since $\frac{\partial \hat{e}}{\partial e_{ns}} = \frac{1}{n} < \frac{\partial \hat{e}}{\partial e_s} = \frac{s}{n} < 1$, for $2 < s < n$. Since each signatory considers the effect of all members on the social norm, it perceives a greater marginal influence on the global emissions norm than a non-signatory. As a result, the self-image effect induces stronger incentives for emission reductions among coalition members, even under symmetric preferences.

Second, the first-order conditions derived from the emissions game allow us to compare the size of the stable coalition under global norm considerations to the case where countries do not value such a norm. Let $B'$, $D'$ and $C'$ denote the derivatives of the benefit, damage, and self-image cost functions, respectively. Assuming countries are homogeneous, the first-order condition for each country $i = s, ns$, is given by, $B' = C' \frac{\partial(\psi i)}{\partial e_i} + D' \frac{\partial E}{\partial e_i}$. The differences between signatories and non-signatories arise from their respective marginal effects on the global emissions norm and total emissions. Specifically, we have $\frac{\partial E}{\partial e_s} = s > \frac{\partial E}{\partial e_{ns}} = 1$ and $\frac{\partial \psi_{ns}}{\partial e_{ns}} = 1 - \frac{1}{n} > \frac{\partial \psi_s}{\partial e_s} = 1 - \frac{s}{n}$, for $s > 1$.

Signatories and non-signatories will choose the same emissions level if the right-hand sides of their first-order conditions are equal. Setting these equal yields, $C' - \frac{1}{n}C' + D' = C' - \frac{s}{n}C' + D's$, which simplifies to,

$$\left(D' - \frac{1}{n}C'\right)(s-1) = 0. \tag{6}$$

By imposing the reasonable assumption that a country's self-image cost from deviating from the global norm does not exceed the marginal environmental damage it imposes on all countries—i.e., $D' - \frac{1}{n}C' > 0$—the condition derived above (6) holds only when $s = 1$.

This reflects a key structural asymmetry: a signatory internalizes the environmental damage it imposes on all $s$ members of the coalition, whereas a non-signatory considers only the damage it suffers individually. As a result, the critical coalition size $s^{\min}$—the value of $s$ at which signatories and non-signatories would choose the same emissions level—is equal to one. This mirrors the outcome in the standard setting where countries do not consider the global emissions norm, see for example (**?**), reinforcing the difficulty of expanding cooperation.

The relationship between the critical coalition size $s^{\min}$ and the size of the stable coalition can be established directly. At the Nash equilibrium—corresponding to $s^{\min} = 1$—signatories and non-signatories select identical emission levels $e_s^*(1) = e_{ns}^*(1) = e_N$, reflecting the fully non-cooperative outcome. At this point, welfare for all countries is at its lowest, denoted by $w_s(1) = w_{ns}(1) = w_N$. As coalition size increases beyond this minimal level—beginning with a coalition of two countries—signatories reduce their emissions relative to non-signatories, as they internalize the external damages imposed on fellow coalition members. In contrast, non-signatories continue to act independently, optimizing without regard to the externalities they generate.

As the coalition grows, aggregate emissions decline, leading to welfare improvements for both members and non-members. Formally, this implies, $\frac{dw_i}{ds} > 0$. Moreover, non-signatories benefit more from the expansion of the coalition, as they enjoy the environmental gains from reduced global emissions without incurring significant reductions in economic benefits. the cost of abatement. Thus, $\frac{dw_{ns}(s)}{ds} > \frac{dw_s(s)}{ds}$.[4] This asymmetry reflects a classic free-rider problem: although expanding the coalition enhances collective welfare, the uneven distribution of costs and benefits weakens individual incentives to participate. As a result, the stable coalition size remains limited, despite the potential for greater global

[4] The proofs are provided in Appendix 1.

gains through broader cooperation.

If a member exits a coalition of size $s$, the welfare of nonsignatories becomes $w_{ns}(s-1)$, exhibiting the same properties as $w_{ns}(s)$ shifted to the right by one. Specifically, the $w_{ns}(s-1)$ function intersects from below the $w_s(s)$ function at a coalition size higher than $s^{\min}$. Adjusting this critical size to the lower integer yields the size of a stable coalition, denoted by $s^*$, since below that point $w_s(s^*) \geq w_{ns}(s^*-1)$, and above it, $w_{ns}(s^*) \geq w_s(s^*+1)$. That is, at $s^*$ both internal and external stability conditions are satisfied.[5]

When working with general functional forms, determining the exact value of the stable coalition size $s^*$ is analytically intractable. The size of the stable coalition depends critically on the slope of the welfare function $w_s(s)$ around the critical threshold $s^{\min}$. Small changes in the curvature of this function can lead to significant variation in $s^*$.

Previous studies using quadratic specifications have shown that the size of the stable coalition does not exceed two members. This finding will be corroborated in the following section through numerical analysis. However, Eckert and Nkuiya (2022) demonstrate that under alternative assumptions—specifically, quadratic benefits combined with exponentially increasing damages—the equilibrium coalition size can be substantially larger, particularly when countries exhibit high sensitivity to environmental damages.

The following Proposition summarizes these insights.

**Proposition 1** *In a simultaneous game, when countries take into account their position relative to a global environmental norm, they choose lower emission levels compared to a scenario where they do not consider the norm. Assuming completely symmetric countries and defining the global norm as the average level of emissions, global norm considerations do not improve the size of stable coalition. Despite lower overall emissions, the incentive structure underlying coalition formation remains unchanged, and the free-rider problem persists.*

In the case that the sensitivity to the global norm differs between signatories and nonsignatories, $C'_{ns} \neq C'_s$, the above results do not change under the reasonable assumption that signatories are more sensitive relative to non-signatories to deviations from the global

[5] Figure 1, illustrating signatories' and nonsignatories' welfare functions for quadratic functions, provides additional clarity.

norm, $C'_s = \frac{\partial C(\psi_s)}{\partial \psi_s} > \frac{\partial C(\psi_{ns})}{\partial \psi_{ns}} = C'_{ns}$. In such case, signatories and non-signatories emission choices coincide if,

$$n(s-1)D' = (n-1)C'_{ns} - (n-s)C'_s. \tag{7}$$

Solving for the critical value of $s$ yields,

$$s^{\min} = 1 - \frac{n-1}{n}\frac{C'_s - C'_{ns}}{D' - \frac{1}{n}C'_s}. \tag{8}$$

Given that $D' - \frac{1}{n}C' > 0$ and $C'_s > C'_{ns}$ the critical value of $s$ is smaller than one.[6] Therefore, the critical coalition size $s^{\min} < 1$ if $C'_s > C'_{ns}$. Consequently, the size of the stable coalition, determined at the right of the point where equilibrium emission choices are equal –corresponding to the intersection of the non-signatories' indirect welfare with that of the signatories– also decreases.

Given that $D' - \frac{1}{n}C' > 0$, larger stable coalitions become possible only if non-signatories are more sensitive to the global norm than signatories, $C'_{ns} > C'_s$.[7] The intuition is as follows: if the difference between $C'_{ns}$ and $C'_s$ is sufficiently large so that sensitivity to the global norm outweighs the effect of damage internalization by coalition members, non-signatories will choose lower emissions than signatories and thus $s^{\min}$ could exceed one. In such scenarios, the stable coalition size lies on the decreasing part of the signatories' indirect utility function. While in this case welfare of both groups improves significantly, we consider it unrealistic and do not explore it further.

## 4 Sequential coalition formation game

In this section, we analyze a leadership setting in which coalition members act as first movers in the determination of emission levels. The coalition jointly selects its emissions level to maximize aggregate welfare, while fully anticipating the strategic responses of non-signatory countries. Non-signatories, by contrast, behave non-cooperatively, choosing their emissions independently to maximize individual welfare, conditional on the coalition's

[6] Both $C'_i$ and $D'$ are functions of their arguments, $\psi_i$ and $E$ respectivelly and thus, in general, depend on $s$. However, evaluated at the equilibrium values of emission choices, $e^*_s = e^*_{ns} = e_N$–which is necessary to specify $s^{\min}$–it follows that $\psi^*_i = 0$ and $E = ne_N$. Consequently, at equilibrium, both $C'_i$ and $D'$ are functions of the model's parameters and are independent of $s$.

[7] Note thought that in such case, the above analysis leading to (8) does not apply since $e_s \neq e_{ns}$ leads to $B(e_{ns}) \neq B(e_s)$.

decision. The model is solved through backward induction: we begin by deriving the non-signatories' best-response functions to any given coalition strategy and then determine the coalition's optimal emissions choice by incorporating these responses into its objective function. As in the simultaneous game, we assume that both signatories and non-signatories take into consideration the global environmental norm when choosing their emissions.

The non-signatories' optimal emissions level, expressed as a function of the coalition's emissions, $e^*_{ns}(e_s)$, is derived from the first-order condition presented in equation (5). As previously discussed, when non-signatories internalize the global emissions norm, they choose lower emissions than they would under a purely self-interested benchmark.

To characterize the strategic interaction between coalition members and non-signatories, we differentiate the non-signatories' first-order condition with respect to $e_s$. This yields the slope of their best-response function. The following lemma establishes that the emissions choice of non-signatories is a strategic substitute to that of the coalition—that is, an increase in the coalition's emissions leads non-signatories to reduce their own emissions.[8]

**Lemma 2** *Non-signatories' emissions choices are strategic substitutes to those of signatories. Moreover, the sensitivity of non-signatories' responses increases when they account for the global environmental norm. Formally, the slope of the non-signatories' best-response function with respect to the signatories' emissions is given by,*

$$\frac{\partial e^*_{ns}(e_s)}{\partial e_s} = -\frac{s\left[n^2 D''(E) - (n-1)C''(\psi_{ns})\right]}{n^2\left((n-s)D''(E) - B''(e_{ns})\right) + s(n-1)C''(\psi_{ns})} < 0. \tag{9}$$

First, observe that the second-order condition for the non-signatories' welfare maximization problem, $(n-s)D''(E) - B''(e_{ns}) > -\frac{s(n-1)}{n^2}C''(\psi_{ns}(s))$, ensures that the non-signatories' best-response function remains downward sloping, i.e., $\frac{\partial e_{ns}}{\partial e_s} < 0$, where the inequality follows from the standard assumptions $B'' < 0$ and $D'' > 0$ and $C'' < 0$. This implies that, under standard concavity assumptions, the introduction of costs from global norm deviations—captured by the curvature of $C(\psi_{ns})$—does not reverse the strategic substitutability between signatories' and non-signatories' emissions decisions. That is, non-signatories will continue to reduce their emissions in response to an increase in the coalition's emissions.

[8] The formal proof is provided in the Appendix.

To evaluate the impact of global norm considerations on the behavior of non-signatories, consider first the benchmark case in which non-signatories do not consider the global emissions norm—that is, $C'' = 0$. In this case, their reaction function simplifies to $\frac{\partial e_{ns}^*(e_s)}{\partial e_s} = -\frac{sD''(E)}{(n-s)D''(E)-B''(e_{ns})} < 0$. In this setting, non-signatories choose their emissions such that the marginal private benefit equals the marginal environmental damage, without regard for normative or reputational effects.

When global norm considerations are introduced—specifically, when $C'' < 0$—the absolute value of the slope of the non-signatories' reaction function increases. This change is driven by two effects: an increase in the numerator due to the additional term $-(n-1)C''(\psi_{ns}) > 0$, and a reduction in the denominator resulting from the added term $s(n-1)C''(\psi_{ns}) < 0$. Consequently, non-signatories' emissions become more sensitive to changes in the coalition's emissions. In effect, stronger ethical or reputational concerns—captured by a more negative $C''$—enhance the coalition's ability to exert leadership power.

Signatories collectively choose their optimal level of emissions $e_s^*(e_{ns}^*(e_s))$, by anticipating the best-response behavior of non-signatories. This is achieved by solving the following maximization problem,

$$\max_{e_s} sW_s = \max_{e_s} \left\{ s\left(B\left(e_s\right) - C\left(\psi_s\left(e_s, e_{ns}^*\left(e_s\right)\right)\right) - D\left(E\right)\right)\right\}, \tag{10}$$

where $\psi_s(e_s, e_{ns}^*(e_s)) = e_s - \hat{e}$, $\hat{e} = \frac{E}{n}$, $E = se_s - (n-s)e_{ns}^*(e_s)$.

The signatories' optimal emissions level, $e_s^*(e_{ns}^*(e_s))$ satisfies the following first-order condition,

$$B' - C' = \left(D' - \frac{1}{n}C'\right)\left[s + (n-s)\frac{\partial e_{ns}^*\left(e_s\right)}{\partial e_s}\right], \tag{11}$$

where, as in the previous section primes denote first order derivatives. This condition captures both the direct effect of a change in a signatory's emissions on its welfare and the indirect effect through the induced change in non-signatories' emissions. The left-hand side of the equation represents the marginal benefit of increased emissions for the coalition, net of the marginal cost associated with falling behind the global emissions norm.

The first bracketed term on the right-hand side captures the marginal environmental damage resulting from an increase in emissions by the coalition, partially offset by a reduction in the self-image cost, as the global emissions norm rises with the coalition's emissions. The second bracketed term reflects the change in aggregate emissions induced by an increase in signatories' emissions. This term incorporates both the direct effect—the immediate increase in the coalition's own emissions—and the indirect effect—the reduction in non-signatories' emissions in response to the coalition's behavior, reflecting the strategic substitutability between the two groups' choices.

As discuissed above, $\frac{\partial}{\partial C''(\psi_{ns})}\left|\frac{\partial^2 e^*_{ns}(e_s)}{\partial e_s}\right| > 0$, indicating that the magnitude of the second bracketed term in (11) decreases as countries become more sensitive to the global norm. Assuming signatories and non-signatories exhibit equal sensitivity, an increase in this sensitivity generates two opposing effects on the emission choices of signatory countries. First, there is a direct effect, whereby heightened sensitivity induces signatories to reduce their emissions. Second, there is an indirect effect, whereby the same sensitivity alters non-signatories' behavior in a way that incentivizes signatories to increase their emissions. For a given coalition size, this indirect effect may outweigh the direct one, leading to a counterintuitive outcome: signatories may increase their emissions as $C''(\psi)$ rises. We demonstrate this possibility through a numerical example in the following Section. Thus, as sensitivity to the global norm increases, signatories may increase their emissions while non-signatories reduce theirs—with the reduction by non-signatories more than offsetting the increase by signatories.

When determining the size of a stable coalition in the leadership setting, it is essential—just as in the simultaneous-move game analyzed in the previous section—to identify the critical coalition size at which the emission choices of signatories and non-signatories coincide. This alignment occurs when both groups face identical first-order conditions for their emission decisions. Under the assumption of homogeneous countries, such that $B'(e_s) - C'(\psi_s) = B'(e_{ns}) - C'(\psi_{ns})$, this condition requires that the right-hand sides of equations (11) and (5) be equal. This leads to the following condition,

$$\left[D' - \frac{C'(\psi_s)}{n}\right]\left[s + (n-s)\frac{\partial e^*_{ns}(e_s)}{\partial e_s}\right] = D' - \frac{C'(\psi_{ns})}{n}, \tag{12}$$

which given the assumption, $C'(\psi_s) = C'(\psi_s) = C'$, simplifies to,

$$\left(D' - \frac{C'}{n}\right)\left[(s-1) + (n-s)\frac{\partial\left(e^*_{ns}(e_s)\right)}{\partial\left(e_s\right)}\right] = 0. \tag{13}$$

It is important to note that, although the resulting condition resembles condition (6) from the simultaneous-move game, it differs in a fundamental way due to the coalition's role as a leader. Specifically, when non-signatories determine their emissions level $e^*_{ns}(e_s)$ according to equation (5) they set $\frac{\partial E}{\partial e_{ns}} = 1$, indicating that they internalize only the marginal environmental damages they personally incur.

In contrast, signatories, when choosing $e^*_s$ under the leadership structure described in equation (11), set $\frac{\partial E}{\partial e_s} = s + (n-s)\frac{\partial(e^*_{ns}(e_s))}{\partial(e_s)}$. This expression reflects that signatories not only account for the marginal damages imposed on all coalition members (captured by the term $s$), but also, as first movers, internalize the effect of their emissions on the behavior of non-signatories. The term $(n-s)\frac{\partial(e^*_{ns}(e_s))}{\partial(e_s)}$ captures how changes in the coalition's emissions influence the aggregate emissions through non-signatories' strategic response.

Assuming once again that the marginal ethical cost of deviating from the global emissions norm does not exceed the marginal environmental damage imposed on all countries—i.e., $nD' - C' > 0$—condition (13) can only be satisfied if the terms within the brackets sum to zero.

The first of these terms, $(s-1)$ , reflects the damage internalization effect discussed earlier. Specifically, when each coalition member selects its emissions level, they internalize not only the marginal environmental damages they experience themselves but also those imposed on the other coalition members, in total $s$ coalition members. In contrast, non-signatories, acting independently, consider only their own marginal damages when making emission decisions.

The second term $(n-s)\frac{\partial(e^*_{ns}(e_s))}{\partial(e_s)}$, captures the reduction in each coalition member's marginal net environmental damage resulting from the strategic adjustment in non-signatories' emissions. Specifically, as signatories increase their emissions, non-signatories respond by reducing their own emissions. This behavioral response mitigates the aggregate environmental damage experienced by the coalition. Using the slope of the non-signatories' reaction function, derived in equation (9), this term becomes, $-(n-s)\frac{s\left[n^2D''-(n-1)C''\right]}{n^2((n-s)D''-B'')+s(n-1)C''}$. This expression defines the strategic interaction effect, which quantifies the adjustment of all $(n-s)$ non-signatories' emissions in response to changes in the coalition's emissions level.

Importantly, the magnitude of this effect increases with the absolute value of $C''$, implying that stronger reputational concerns among non-signatories amplify the coalition's ability to influence global emissions through its leadership. This is the only difference from the standard leadership model without global norm considerations (see for example, **?**).

The emission choices of signatories and non-signatories align when the damage internalization effect balances the strategic interaction effect. By substituting the expression $\frac{\partial(e^*_{ns}(e_s))}{\partial(e_s)}$ from (9) into condition $(s-1)+(n-s)\frac{\partial(e^*_{ns}(e_s))}{\partial(e_s)}=0$, we derive the critical value of $s$ at which the equilibrium emissions of non-signatories and signatories are equal, i.e., $e^*_{ns}=e^*_s$,

$$s^{\min}=\frac{nD''-B''}{D''-B''+(\frac{n-1}{n})^2C''}. \tag{14}$$

The critical coalition size $s^{\min}$ derived here differs from the standard case—see Proposition 1 in Diamantoudi and Sartzetakis (2015)—due to the presence of the additional term $(\frac{n-1}{n})^2C''<0$ in the denominator. As the absolute value of $C''$ increases (i.e., as global norm considerations become more salient), the denominator decreases, thereby increasing the value of $s^{\min}$.

This result implies that stronger norm concerns among non-signatories enhance the coalition's leadership effectiveness, making participation in emission reduction efforts more attractive. As a consequence, the critical coalition size required for emissions alignment increases, supporting the formation of larger stable coalitions. In essence, when the moral cost of deviating from the global norm rises, so does the incentive for countries to join and remain within the cooperative agreement.

The key insight is summarized in the following proposition.

**Proposition 3** *When the coalition assumes a leadership role in the emissions decision stage, the size of the stable coalition increases if countries internalize the moral costs of falling behind the global environmental norm. Furthermore, the internalization of such normative considerations results in lower aggregate emissions compared to a scenario in which countries act solely out of economic self-interest. However, for a given coalition size, an increase in sensitivity to the global norm can lead to higher emissions by signatories. This increase is offset by a larger reduction in emissions by non-signatories, leading to a net decline in global emissions.*

If signatories and non-signatories differ in their sensitivity to deviations from the global emissions norm, condition (13) becomes,

$$\left(D'-\frac{1}{n}C'_s\right)\left[(s-1)+(n-s)\frac{\partial\left(e^*_{ns}(e_s)\right)}{\partial\left(e_s\right)}\right]=\frac{1}{n}\left(C'_s-C'_{ns}\right). \tag{15}$$

Define $\frac{(C'_s - C'_{ns})}{(nD' - C'_s)} \equiv \Xi$, where for $nD' - C'_s > 0$ and $C'_s > C'_{ns}$, $0 < \Xi < 1$. Solving the above equation for the critical value of $s$,yields,[9]

$$s^{\min} = \frac{n^2 (1 - \Xi) (nD'' - B'')}{n^2 ((1 - \Xi) D'' - B'') + (n - 1) (n - (1 - \Xi)) C''_{ns}}. \tag{16}$$

The value of $s^{\min}$ is decreasing in $\Xi$.[10] This implies that as signatories' sensitivity to the global norm increases relative to that of non-signatories, the critical value $s^{\min}$ declines. Assuming $nD' - C'_s > 0$, $s^{\min}$ also decreases with the difference $C'_s > C'_{ns}$, because the strategic effect weakens. Highly motivated signatories, influenced more strongly by the global norm, adopt lower emission levels. As a result, their ability to exert leadership influence diminishes.

It is evident that if $C'_s < C'_{ns}$ and $nD' - C'_s > 0$, the size of the coalition would increase, as the strategic interaction effect would grow. The coalition size would increase, as the strategic interaction effect becomes stronger. However, as discussed in the previous section, this assumption is not realistic. Consequently, we do not further consider this scenario—just as we exclude the case where $nD' - C'_s \leq 0$.

# 5 Illustration of the results using quadratic functions

In this section, we adopt quadratic benefit functions, which are commonly used in the literature. Specifically, we define the benefit to country $i$ as, $B(e_i) = ab(e_i) - \frac{b}{2}(e_i)^2$, and the cost of deviating from the global norm as $C(\psi_i) = a\delta(\psi_i) - \frac{\delta}{2}(\psi_i)^2$, where $a > 0$, $b > 0$ and $\delta > 0$. The damages from global emissions are given by $D(E) = \frac{c}{2}E^2 = \frac{c}{2}\left(\sum_{i=1}^n e_i\right)^2$, where $c > 0$. The parameter $\delta$ captures the weight assigned to global norm considerations. Its value relative to $b$ reflects the importance of normative concerns compared to economic welfare, and we assume $b > \delta$. When $\delta = 0$ the social welfare function of country $i$ reduces to that of the baseline IEAs model, focusing solely on the economic benefits and environmental damages associated with emissions.

By substituting these specific functional forms into expression (1), we define the welfare

[9] Recall from the previous Section that both $C'_i$ and $D'$ evaluated at the equilibrium where, $e^*_s = e^*_{ns} = e_N$ are functions of the model's parameters and are independent of $s$.

[10] The derivative is, $\frac{\partial s^{\min}}{\partial \Xi} = -\frac{n^2(nD'' - B'')(-n^2 B'' + n(n-1)C''_{ns})}{\left[n^2((1-\Xi)D'' - B'') + (n-1)(n-(1-\Xi))C''_{ns}\right]^2} < 0$, since $-nB'' > -(n-1)C''_{ns}$.

of country $i$ as follows,

$$W\left(e_i,\psi_i,E\right)=b\left(ae_i-\frac{1}{2}e_i^2\right)-\delta\left(a\psi_i-\frac{1}{2}\psi_i^2\right)-\frac{c}{2}E^2. \tag{17}$$

### 5.1 Simultaneous choice of emissions

We examine first the case in which signatories and non-signatories choose emissions simultaneously. Non-signatory countries choose their level of emissions, $e_{ns}$, to maximize their individual welfare. Solving the first-order condition yields the non-signatories' reaction function,

$$e_{ns}^*\left(e_s\right)=a\frac{\Phi_{ns}}{\Omega_{ns}}-\frac{\Psi_{ns}}{\Omega_{ns}}se_s, \tag{18}$$

where, $\Phi_{ns}=b-v_{ns}\delta_{ns}>0$, $\Psi_{ns}=c+\frac{v_{ns}}{n}\delta_{ns}>0$, $\Omega_{ns}=\Phi_{ns}+(n-s)\Psi_{ns}=b+(n-s)\,c-\frac{s}{n}v_{ns}\delta_{ns}>0$ and $0<v_{ns}=\frac{n-1}{n}<1$, for $n>2$. The slope of the non-signatories' reaction function with respect to the signatories' emissions is negative, $\frac{\partial e_{ns}^*(e_s)}{\partial e_s}<0$, since the condition $\Omega_{ns}>0$, or $b+(n-s)\,c>\frac{s}{n}v_{ns}\delta_{ns}$ holds.

Signatory countries maximize their collective welfare. Solving the first-order condition of their optimization problem yields the signatories' reaction function,

$$e_s^*\left(e_{ns}\right)=a\frac{\Phi_s}{\Omega_s}-\frac{\Psi_s}{\Omega_s}(n-s)e_{ns}, \tag{19}$$

where, $\Phi_s=b-v_s\delta_s>0$, $\Psi_s=sc+\frac{v_s}{n}\delta_s>0$, $\Omega_s=\Phi_s+s\Psi_s=b+s^2c-v_s^2\delta_s>0$ and $0\leq v_s=\frac{n-s}{n}\leq 1$. We confirm that the slope of signatories' reaction function is also negative, $\frac{\partial e_n^*(e_{ns})}{\partial e_{ns}}<0$, since the condition $\Omega_s>0$ holds.

Solving the system of reaction functions in (19) and (18) yields the optimal emission levels, $e_s^*$ and $e_{ns}^*$, expressed as functions of the model parameters,

$$e_s^{c*}=a\frac{\Phi_s\Omega_{ns}-(n-s)\Psi_s\Phi_{ns}}{\Omega_s\Omega_{ns}-s(n-s)\Psi_s\Psi_{ns}}, \tag{20}$$

$$e_{ns}^{c*}=a\frac{\Phi_{ns}\Omega_s-s\Psi_{ns}\Phi_s}{\Omega_s\Omega_{ns}-s(n-s)\Psi_s\Psi_{ns}}. \tag{21}$$

Using the quadratic specification of the model, the critical coalition size, $s^{\min}$–defined as the coalition size at which signatories and non-signatories emit the same amount, $e_s^*\left(s^{\min}\right)=e_{ns}^*\left(s^{\min}\right)=e_N$–is given by,

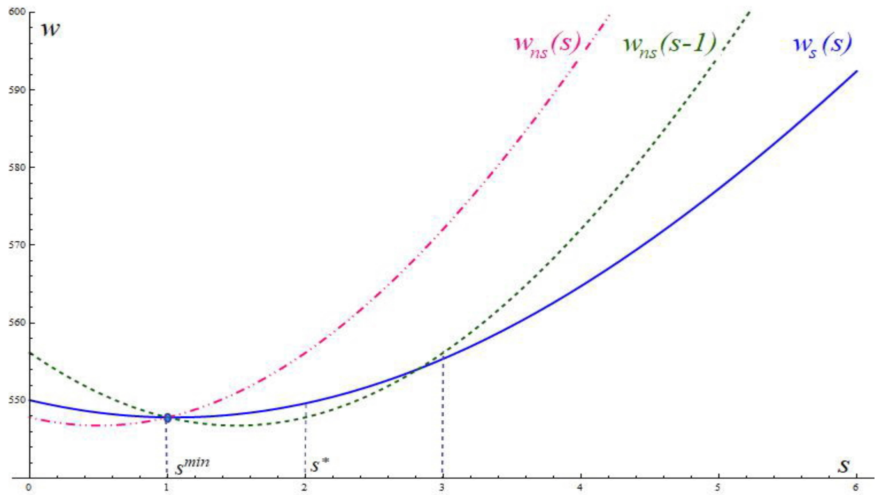
$$s^{\min} = 1 - \frac{n-1}{n} \frac{a\left(\delta_s - \delta_{ns}\right)}{cne_N - \frac{1}{n}\delta_s a}, \tag{22}$$

where $e_N$ is the emission level at the Nash equilibrium (i.e., when all countries act non-cooperatively), and is defined as $e_N = a\frac{b - v_{ns}\delta_{ns}}{b + nc - \delta_{ns}}$. Therefore, when $\delta_s = \delta_{ns}$ the critical coalition size simplifies to, $s^{\min} = 1$, which aligns with the standard result in the case of a simultaneous-move game. However, when $\delta_s > \delta_{ns}$, the adjustment term $0 < v_{ns}\frac{a(\delta_s - \delta_{ns})}{cne_N - \frac{1}{n}\delta_s a} < 1$, is strictly positive and less than one. This leads to a smaller critical coalition size, $s^{\min} < 1$, indicating that the size of stable agreement is smaller when signatories place greater weight on global norm considerations compared to non-signatories.

In the following, we illustrate the results by presenting numerical examples across a range of parameter values. To achieve this, we substitute the equilibrium emission levels for signatories and non-signatories, as derived from (20) and (21), into the welfare function outlined in (17). This allows us to derive the indirect welfare functions for both signatories and non-signatories. Next, we apply the internal and external stability conditions, as defined in (2) and (3), to determine the size of stable coalitions. We begin by presenting a graphical representation of how the size of the stable coalition is determined, followed by the results for a comprehensive set of admissible parameter values.

Figure 1 illustrates the indirect welfare of countries as a function of the coalition size $s$, using the following parameter values: $n = 10$, $a = 10$, $b = 20$, $c = 0.1$ and $\delta_s = \delta_{ns} = \delta = 0.1$. The solid blue curve represents the indirect welfare of signatories, $w_s(s)$, the red dashed-dotted curve shows the indirect welfare of non-signatories, $w_{ns}(s)$, and the green dashed curve represents the indirect welfare of non-signatories shifted by one, $w_{ns}(s-1)$. The figure focuses on the range of $s$ values where interesting intersections occur, as for higher values of $s$, there are no further intersections.

A coalition of two countries is both internally and externally stable, as illustrated in the graph. The actual welfare values confirm that the stable coalition size is $s^* = 2$, satisfying both the internal (2) and external (3) stability conditions: $w_s\left(s=2\right) = 549.67 > w_{ns}\left(s=1\right) = 547.86$, and $w_s\left(s=3\right) = 555.40 < w_{ns}\left(s=2\right) = 556.17$. This indicates that as soon as the smallest coalition forms, free-riding incentives dominate, leading to non-signatories consistently enjoying higher welfare than signatories in larger coalitions. Therefore, the only stable agreement involves a coalition of two members. It is worth noting that at $s^* = 2$, the welfare gain over the Nash baseline $w_N = 547.86$ is minimal, 1 new




1.jpg

Figure 1: Stable coalition size in the simultaneous game: $n = 10$, $b = 20$, $a = 10$, $c = 0.1$ and $\delta = 0.1$.

relative to the potential of the grand coalition $w_s\,(s = 10) = 666.67$.

To verify Proposition 1, we conduct a comparative static analysis with respect to countries' sensitivity to the global norm, represented by the parameter $\delta$. Table 1 presents four scenarios with $\delta$ values ranging from 0 to 0.9—the maximum value satisfying $\delta < b$ and $\delta < nc$. All other parameters are held constant, matching those used in Figure 1.

The simulation results in Table 1 confirm that emissions from both signatories and non-signatories decrease as sensitivity to the global norm ($\delta$) increases. Additionally, the size of the stable coalition remains constant at $s^* = 2$, across all admissible values of $\delta$. Although the level of cooperation does not increase, the reduction in emissions driven by the global norm enhances the welfare of both signatories and non-signatories, thereby raising aggregate welfare above the Nash equilibrium level.

In cases that signatories sensitivity to the global norm differs from that of the non-signatories we verify first, that if $\delta_s > \delta_{ns}$ emissions reduction is smaller relative to the case that both have high level of sensitivity to the global norm. For example, if $\delta_s = 0.5$ and $\delta_{ns} = 0$, signatories reduce their emissions slightly more $e_s = 8.85$ relative to the

| Variable | Value | | | |
|---|---|---|---|---|
| $\delta$ | 0 | 0.10 | 0.50 | 0.90 |
| *Signatories* | | | | |
| $\mathbf{s}^{c*}$ | 2 | 2 | 2 | 2 |
| $e_s^{c*}$ | 9.06 | 9.02 | 8.87 | 8.72 |
| $w_s^{c*}$ | 546.10 | 549.67 | 563.40 | 576.28 |
| $\psi_s^{c*}$ | $-0.38$ | $-0.37$ | $-0.36$ | $-0.339$ |
| *Non-signatories* | | | | |
| $e_{ns}^{c*}$ | 9.53 | 9.49 | 9.32 | 9.15 |
| $w_{ns}^{c*}$ | 552.78 | 556.17 | 569.22 | 581.46 |
| $\psi_{ns}^{c*}$ | 0.094 | 0.093 | 0.089 | 0.084 |
| *Aggregate emissions* | | | | |
| $E^{c*}$ | 94.34 | 93.92 | 92.27 | 90.61 |
| | | | | |
| *Total welfare* | | | | |
| $\sum_{i=1}^{n} w_i^{c*}$ | 5514 | 5549 | 5681 | 5804 |

Table 1: Numerical simulations of the simultaneous moves game. n=10, a=10, b=20, c=0.1 and increasing values of delta.

respective value in column four of Table 1, to compensate for the higher emission level of non-signatories $e_{ns} = 9.53$. As a result aggregate emissions are higher and welfare of both signatories and non-signatories–and thus, aggregate welfare $\sum_{i=1}^{n} w_i = 5549 < 5681$–are lower relative to the case in which $\delta_s = \delta_{ns} = 0.5$.

Second, in the unlikely case that non-signatories are more sensitive to the norm than signatories, $\delta_{ns} > \delta_s$, we verify that the coalition size can increase. For example, if $\delta_s = 0$ and $\delta_{ns} = 0.9$, a coalition of three countries is stable, $s^* = 3$, since both the internal and external stability conditions are satisfied: $w_s\,(s = 3) = 576.71 > w_{ns}\,(s = 2) = 575.84$, and $w_s\,(s = 4) = 581.99 < w_{ns}\,(s = 3) = 586.39$. Substituting the parameters ' values into (19) the critical value of $s$ is, $s^{\min} = 1.85$ and the intersection of $w_{ns}(s-1)$ with $w_s(1)$ is above $s = 3$.

## 5.2 Coalition as a leader in the emissions choice game

We now examine the case in which the coalition, acting as a leader, chooses its emissions to maximize the aggregate welfare of its members, accounting for the reaction of non-signatories as described in equation (18),

$$e_{ns}^{*}\left(e_{s}\right)=a\frac{\Phi_{ns}}{\Omega_{ns}}-\frac{\Psi_{ns}}{\Omega_{ns}}se_{s}. \tag{23}$$

Note that $\frac{\partial e_{ns}^{*}(e_s)}{\partial e_s}<0$, and that, $\frac{\partial}{\partial \delta_{ns}}\left|\frac{\partial^2 e_{ns}^{*}(e_s)}{\partial e_s}\right| = s\frac{n^2(n-1)(b+nc)}{[n^2(b+(n-s)c)-s(n-1)\delta_{ns}]^2}>0$, which confirms Lemma 1: the greater the sensitivity of non-signatories to the global environmental norm, the more responsive they are to changes in the emissions of signatories.

From the first-order condition of signatories, we derive their equilibrium emissions level,

$$e_{s}^{l*}=a\frac{\Phi_{s}^{l}\Omega_{ns}-(n-s)\Psi_{s}^{l}\Phi_{ns}}{\Omega_{s}^{l}\Omega_{ns}-s(n-s)\Psi_{s}^{l}\Psi_{ns}}, \tag{24}$$

where, $\Phi_{s}^{l}=b-v_{s}^{l}\delta_{s}>0$, $\Psi_{s}^{l}=\gamma c+\frac{v_{s}^{l}}{n}\delta_{s}>0$, $\Omega_{s}=\Phi_{s}^{l}+s\Psi_{s}^{l}$ and $v_{s}^{l}=\frac{n-\gamma}{n}>v_{s}$, with $\gamma=\frac{\partial E}{\partial e_s}=s-(n-s)s\frac{\Psi_{ns}}{\Omega_{ns}}<s$. Substituting equation (24) into the non-signatories' reaction function (18), we obtain their equilibrium emissions as a function of the model parameters,

$$e_{ns}^{l*}=a\frac{\Phi_{ns}\Omega_{s}^{l}-s\Psi_{ns}\Phi_{s}^{l}}{\Omega_{s}^{l}\Omega_{ns}-s(n-s)\Psi_{s}^{l}\Psi_{ns}}. \tag{25}$$

If countries are unresponsive to the global environmental norm (i.e., $\delta_s=\delta_{ns}=\delta=0$), both signatories' and non-signatories' emissions reduce to the standard model solutions, as described by equations (4) and (5) in Diamantoudi and Sartzetakis (2006).

We can derive the critical coalition size, $s^{\min}$, by equating the emissions of signatories and non-signatories: $e_{s}^{*}\left(s^{\min}\right)=e_{ns}^{*}\left(s^{\min}\right)$ using (24) and (25). Due to the complexity of the resulting expression when $\delta_s\neq\delta_{ns}$, we report only the case that both signatories and nonsignatories have the same responsiveness to the global environmental, $\delta_s=\delta_{ns}=\delta$, which yields,

$$s^{\min}=\frac{cn+b}{c+b-(\frac{n-1}{n})^{2}\delta}. \tag{26}$$

This expression is consistent with the general formulation of $s^{\min}$ in equation (14). Furthermore, when $\delta=0$, it collapses to the result reported in Diamantoudi and Sartzetakis (2006).

We derive the indirect welfare functions for signatories and non-signatories by substituting their respective equilibrium emissions, from equations (24) and (25), into the welfare function defined in (17).

As in the simultaneous moves game, it is analytically intractable to derive the stable coalition size using the internal and external stability conditions from (2) and (3). We therefore rely on numerical simulations, first presenting a graphical illustration of how the stable coalition size is determined, followed by results over a broad set of admissible parameter values.

Figure 2 plots the indirect welfare of countries as a function of coalition size $s$, using the following parameter values: $n = 10$, $a = 10$, $b = 6$, $c = 0.39999$ and $\delta_s = \delta_{ns} = \delta = 3.7$. As in Figure 1, the solid blue curve represents the indirect welfare of signatories, $w_s(s)$, the red dashed-dotted curve shows the indirect welfare of non-signatories, $w_{ns}(s)$, and the green dashed curve represents the indirect welfare of non-signatories shifted by one, $w_{ns}(s-1)$. The figure focuses on the range of $s$ values where relevant intersections occur. For larger $s$, no further intersections are observed. Relative to Figure 1, the parameters $b$ and $c$ have been modified to match the example in Diamantoudi and Sartzetakis (2006). In that baseline case with $\delta = 0$ the stable coalition size is $s^*(\delta = 0) = 3$.

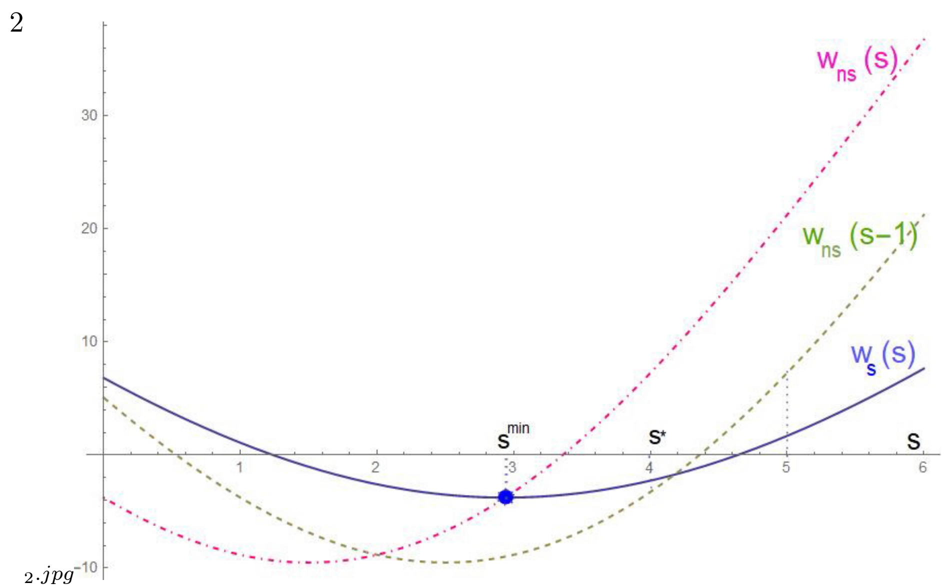



Figure 2: Stable coalition size in the leadership game: $n = 10$, $b = 6$, $a = 10$, $c = 0.39999$ and $\delta = 3.7$.

As countries' sensitivity to the global emissions norm increases from zero to 3.7, the critical coalition size rises from $s^{\min}(\delta = 0) = 1.56$ to $s^{\min}(\delta = 3.7) = 2.93$. Figure

2 illustrates that for $\delta = 3.7$, a coalition of four countries is stable, i.e., $s^*(\delta = 3.7) = 4$. Specifically, the coalition is internally stable, since $w_s\,(s = 4) > w_{ns}\,(s = 3)$, and externally stable, as $w_s\,(s = 5) < w_{ns}\,(s = 4)$. This numerical example supports the first result of Proposition 2, namely that climate agreements are more likely to succeed as countries become more sensitive to the global environmental norm.

Table 2 presents simulation results for increasing values of $\delta$. Notably, for values of $\delta$ up to 2, the size of the stable coalition remains constant at $s^* = 3$. Over this range, emissions by signatories ($e_s^{l*}$) increase slightly. This is consistent with our earlier result: as $\delta$ increases, the absolute slope of the non-signatories' reaction function increases, and for a given coalition size $s$, the marginal impact of a signatory's emissions on total emissions, $\frac{\partial E}{\partial e_s}$ decreases. This confirms the final part of Proposition 2, which predicts that under certain conditions, higher sensitivity to norms can induce signatories to increase emissions. However, once $\delta$ becomes large enough for a four-member coalition to be stable—as seen in the final column of Table 2—signatories' emissions begin to decline.

| Variable | Value | | | | |
|---|---|---|---|---|---|
| $\delta$ | 0 | 0.5 | 1 | 2 | 3.7 |
| *Signatories* | | | | | |
| $s^*$ | 3 | 3 | 3 | 3 | 4 |
| $e_s^{l*}$ | 2.73 | 2.75 | 2.77 | 2.78 | 1.83 |
| $w_s^{l*}$ | $-427$ | $-347$ | $-274$ | $-148$ | $-2.26$ |
| $\psi_s^{l*}$ | $-2.60$ | $-2.26$ | $-1.91$ | $-1.20$ | $-0.83$ |
| *Non-signatories* | | | | | |
| $e_{ns}^{l*}$ | 6.45 | 5.98 | 5.50 | 4.50 | 3.01 |
| $w_{ns}^{l*}$ | $-306$ | $-256$ | $-207$ | $-118$ | 7.26 |
| $\psi_{ns}^{l*}$ | 1.1 | 1.0 | 0.8 | 0.5 | 0.36 |
| *Aggregate Emissions* | | | | | |
| $E^{l*}$ | 53.3 | 50.1 | 46.8 | 39.8 | 26.62 |
| | | | | | |
| *Total Welfare* | | | | | |
| $\sum_{i=1}^{n} w_i^{l*}$ | $-3425$ | $-2832$ | $-2272$ | $-1273$ | 44.02 |

Table 2: Numerical simulation of the leadersip game, n=10, a=10, b=6, c=0.39999 and increasing values of delta.

An increase in countries' sensitivity to the global emissions norm significantly enhances environmental quality by sharply reducing aggregate emissions. This decline in emissions lowers environmental damages and, in turn, leads to substantial improvements in the welfare of both signatories and non-signatories. The welfare gains reflect a shift in priorities: countries accept lower economic returns in exchange for higher environmental standards. This also reduces the benefits of free-riding, thereby supporting the formation of larger, more effective coalitions.

When the sensitivity to the global norm differs between signatories and non-signatories, two key outcomes emerge. First, if signatories are more sensitive than non-signatories (i.e., $\delta_s > \delta_{ns}$) overall emission reductions are smaller than in cases where both groups exhibit a high level of sensitivity. For example, when $\delta_s = 1$ and $\delta_{ns} = 0$, both signatories and non-signatories emit significantly more—$e_s^{l*} = 4.09 > 2.77$ and $e_{ns}^{l*} = 7.24 > 5.50$–compared to the case where $\delta_s = \delta_{ns} = 1$ as reported in column four of Table 2. As a result, aggregate emissions rise to $E^{l*} = 66.1 > 46.8$ and both signatories' and non-signatories' welfare levels decline. Total welfare in this asymmetric case is $\sum_{i=1}^{n} w_i^{l*} = -2803$, significantly lower than the $-2272$ observed under symmetric sensitivity $\delta_s = \delta_{ns} = 1$.

Second, in the less likely case where non-signatories are more sensitive to the global norm than signatories (i.e., $\delta_{ns} > \delta_s$), the coalition size can expand even further. For instance, if $\delta_s = 0$ and $\delta_{ns} = 3.7$, the critical coalition size increases to $s^{\min} = 7.24$, and a coalition of eight countries becomes stable ($s^* = 8$). In this scenario, signatories emit $e_s^{l*} = 1.99$ and enjoy welfare $w_s^{l*} = -0.02$, while non-signatories emit $e_{ns}^{l*} = 3.62$ and achieve significantly higher welfare, $w_{ns}^{l*} = 25.04$. Total welfare is $\sum_{i=1}^{n} w_i^{l*} = 49.95$, demonstrating that stronger normative pressure on non-signatories can lead to both larger coalitions and improved collective outcomes, even when signatories themselves are less norm-sensitive.

# 6 Conclusions

his paper explores how moral/ethical considerations affect countries' emission decisions and the formation of stable coalitions. In contrast to standard models that focus solely on net economic gains (benefits from one's own emissions minus the damages from global emissions), we introduce a global environmental norm. Countries experience additional moral costs when they exceed this norm and moral benefits when they perform better than the norm. Concretely, we assume that the global norm is the average level of global emissions. Our framework thus captures the idea that countries value their global environmental standing, reflecting the expectations and environmental awareness of their citizens.

We examine both simultaneous and sequential coalition formation games. Our findings show that incorporating ethical or moral motivations generally reduces aggregate emissions and raises overall welfare, but does not always increase the size of the stable coalition. In the simultaneous-move setting, when both signatories and non-signatories share the same degree of ethical considerations, emissions decline across all countries, yielding notable welfare gains - yet the depth of cooperation does not improve significantly. On the other hand, when coalition members possess stronger ethical motivations than non-members, the stable coalition becomes smaller, as free-riders prefer to exploit the signatories' additional abatement efforts.

Under a leadership structure, if signatories and non-signatories have equal ethical motivations, emissions fall and the size of the coalition increases relative to the scenario without moral considerations. However, if signatories place greater weight on moral aspects than free-riders do, the coalition often shrinks, and significant disparities in ethical motivation can even increase total emissions and reduce overall welfare. Thus, unless non-signatories raise their ethical considerations, improvements stemming from highly motivated signatories alone remain limited.

These results highlight that strong, uniform ethical motivations across all countries are more effective at curtailing emissions and increasing total welfare than situations in which only a fraction of countries consider ethical imperatives. To enhance global climate outcomes, it is crucial not only to strengthen the ethical commitment of signatory countries but also to persuade non-signatories to align more closely with a shared environmental norm. In doing so, countries can realize mutual benefits, particularly under a leadership arrangement, by jointly lowering emissions and delivering greater welfare gains. These findings underscore the importance of ethical considerations and public pressure in driving the adoption of rigorous environmental policies globally.

Future research could explore alternative approaches to defining the global environmental norm, such as basing it on historical emissions or sector-specific standards, rather than the average of total emissions. Moreover, examining asymmetries in how individual countries derive economic benefits from emissions, alongside varying degrees of ethical incentives, would deepen understanding of how heterogeneous preferences influence coalition formation, leadership strategies, and overall welfare outcomes.

## 7 References


Block, S., Emerson, J. W., Esty, D. C., de Sherbinin, A., Wendling, Z. A., et al. (2024). 2024 Environmental Performance Index. New Haven, CT: Yale Center for Environmental Law & Policy, Available at: epi.yale.edu

Breton, M. & Sbragia, L. (2023). 'Self-image and the stability of international environmental agreements.' *Ecological Economics,* **211**, 107869.

Burck, J., Uhlich, T., Bals, C., Hohne, N. & Nascrimiento, L. (2023). 'Climate change performance index 2024'. Available at: URL: https://ccpi.org/download/climate-change-performance-index-2024/

Climate Change Performance Index (CCPI) (2025). Accessed at: https://ccpi.org/countries/

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

Daube, M. & Ulph, D. (2016). 'Moral behaviour, altruism and environmental policy.' *Environmental and Resource Economics*, **63(2)**, 505–522.

Dernbach, J. C. & Brown, D. A. (2008). 'The ethical responsibility to reduce energy consumption.' *Hofstra Law Review,* **37**, 985.

Diamantoudi, E. & Sartzetakis, E. S. (2006). 'Stable international environmental agreements: An analytical approach.' *Journal of Public Economic Theory*, **8(2)**, 247–263.

Diamantoudi, E. & Sartzetakis, E. S. (2015). 'International environmental agreements: coordinated action under foresight.' *Economic Theory*, **59**, 527–546.

Eckert, H. & Nkuiya, B. (2022). 'Damage sensitivity and stability in international environmental agreements.' *Oxford Economic Papers*, **74(4)**, 1063–1076.

Eichner, T. & Pethig, R. (2024a). 'How important are IEAs for mitigation if countries are of the homo moralis type?' CESifo Working Paper No. 11040.

Eichner, T. & Pethig, R. (2024b). 'International environmental agreements when countries behave morally.' *Journal of Environmental Economics and Management*, **125**, 102955.

Finus, M., Furini, F. & Rohrer, A. V. (2024). 'The Stackelberg vs. Nash-cournot folk-theorem in international environmental agreements.' *Economics Letters*, **234**, 111481.

Gróning, C. & Peters, W. (2010). 'Can justice and fairness enlarge international environmental agreements?' *Games*, **1(2)**, 137–158.

McEvoy, D. M., McGinty, M., Cherry, T. L. & Kroll, S. (2024). 'International climate agreements under the threat of solar geoengineering.' *Journal of the Association of*

*Environmental and Resource Economists*, **11(4)**, 853–886.

Organisation for Economic Co-operation and Development (OECD) (2025). Environmental Country Reviews (EPCR). Available at: URL: https://www.oecd.org/en/topics/environmental-country-reviews.html

Peiró-Signes Á., R. Cervelló-Royo, M. Segarra-Oña (2022). 'Can a country's environmental sustainability exert in‡uence on its economic and financial situation? The relationship between environmental performance indicators and country risk.' *Journal of Cleaner Production*, **375**, 134121.

Rashid, R. (2023). Ethical Challenges of Greenhouse Gas Emissions, Technical report, Center for Open Science.

Sacco, A. (2024). 'Altruistic behavior and international environmental agreements: a differential game approach.' *Decisions in Economics and Finance,* https://doi.org/10.1007/s10203-024-00461-1.

Sachs, J., Lafortune, G. & Fuller, G. (2024). The SDGs and the UN Summit of the Future. Sustainable Development Report 2024. Paris: SDSN, Dublin: Dublin University Press. 10.25546/108572. Available at: https://sdgtransformationcenter.org/reports/sustainable-development-report-2024.

Schopf, M. (2023). 'Self-enforcing international environmental agreements and altruistic preferences.' *Environmental and Resource Economics*, **87**, 2309–2359.

Van der Pol, T., Weikard, H.-P. & van Ierland, E. (2012). 'Can altruism stabilise international climate agreements?' *Ecological Economics*, **81**, 112–120.

World Population Review (2025). Environmental Performance Index by Country. Available at: URL: https://worldpopulationreview.com/country-rankings/environmental performance-index-by-country

# 8 Appendix

**Claim 4** *The slope of signatories indirect utility function with respect to s becomes zero if* $nD' = C_s'$*, even when* $e_s \neq e_{ns}$*.*

**Proof.** Below we present the proof of Claim 4.

Tottaly differentiating signatories' indirect welfare function, $w_s(e_s^*(s)) \equiv B\left(e_s^*(s)\right) -$

$D\left(E^{*}(s)\right)-C_{s}(\psi_{s}^{*}\left(e_{s}^{*}(s),E^{*}(s)\right)$, with respect to the coalition size yields,

$$\begin{aligned} \frac{dw_s(s)}{ds} &= \left[B'(e_s^*(s)) - C_s'\left(\psi_s^*\right)\right]\frac{de_s^*(s)}{ds} - \\ &- \left[D'(E^*(s)) - \frac{C_s'\left(\psi_s^*\right)}{n}\right]\left[e_s^*(s) - e_{ns}^*(s) + s\frac{de_s^*(s)}{ds} + (n-s)\frac{de_{ns}^*(s)}{ds}\right], \end{aligned} \tag{27}$$

which can be rewritten as follows,

$$\begin{aligned} \frac{dw_s(s)}{ds} &= \frac{de_s^*(s)}{ds}\left[B'(e_s^*(s)) - C_s'\left(\psi_s^*\right) - s\left(D'(E^*(s)) - \frac{C_s'\left(\psi_s^*\right)}{n}\right)\right] \\ &- \left[D'(E^*(s)) - \frac{C_s'\left(\psi_s^*\right)}{n}\right]\left[\left[e_s^*(s) - e_{ns}^*(s)\right] + (n-s)\frac{de_{ns}^*(s)}{ds}\right]. \end{aligned} \tag{28}$$

But we know that in equilibrium,

$$B'(e_s^*(s)) - C_s'\left(\psi_s^*\right) \equiv s\left(D'(E^*(s)) - \frac{C_s'\left(\psi_s^*\right)}{n}\right). \tag{29}$$

Hence,

$$\frac{dw_s(s)}{ds} = -\left[D'(E^*(s)) - \frac{C_s'\left(\psi_s^*\right)}{n}\right]\left[\left[e_s^*(s) - e_{ns}^*(s)\right] + (n-s)\frac{de_{ns}^*(s)}{ds}\right]. \tag{30}$$

Differentiating both sides of non-signatories' first order condition with respect to $s$ yields,

$$X_{ns}\frac{de_{ns}^*(s)}{ds} = \Phi_{ns}\left[e_s^*(s) + s\frac{de_s^*(s)}{ds} - e_{ns}^*(s) + (n-s)\frac{de_{ns}^*(s)}{dz}\right].$$

where, $X_{ns} = B''(e_{ns}^*(s)) - \frac{n-1}{n}C_{ns}''\left(\psi_{ns}^*\right) < 0$, and $\Phi_{ns} = D''(E^*(s)) - \frac{n-1}{n^2}C_{ns}''\left(\psi_{ns}^*\right) > 0$. Therefore,

$$\frac{de_{ns}^*(s)}{ds} = \frac{\Phi_{ns}\left[\left[e_s^*(s) - e_{ns}^*(s)\right] + s\frac{de_s^*(s)}{ds}\right]}{\Omega_{ns}}, \tag{31}$$

where, $\Omega_{ns} = X_{ns} - (n-s)\Phi_{ns} < 0$.

In a similar manner, differentiating both sides of signatories' first order condition with respect to $s$ yields,

$$\frac{de_s^*(s)}{ds} = \frac{s\Phi_s\left[\left[e_s^*(z) - e_{ns}^*(z)\right] + (n-s)\frac{de_{ns}^*(s)}{ds}\right] + \left[D'(E^*(s)) - \frac{C_s'(\psi_s^*)}{n}\right]}{\Omega_s}. \tag{32}$$

where, $X_s = B''(e_s^*(s)) - \frac{n-s}{n}C_s''\left(\psi_s^*\right) < 0$, $\Phi_s = D''(E^*(z)) - \frac{n-s}{n^2}C_s''\left(\psi_s^*\right) > 0$ and $\Omega_s =$

$X_s - s^2\Phi_s < 0$.

Solving the above system yields the slopes of $e^*_{ns}(s)$ and $e^*_s(s)$ with respect to $s$,

$$\frac{de^*_s(s)}{ds} = \frac{\Phi_s\left[\Omega_{ns} + (n-s)\Phi_{ns}\right]}{\Omega_s\Omega_{ns} - s(n-s)\Phi_s\Phi_{ns}}\left[e^*_s(z) - e^*_{ns}(z)\right]. \tag{33}$$

$$\frac{de^*_{ns}(s)}{ds} = \frac{\Phi_{ns}\left(\Omega_s + s\Phi_s\right)}{\Omega_s\Omega_{ns} - s(n-s)\Phi_s\Phi_{ns}}\left[e^*_s(z) - e^*_{ns}(z)\right]. \tag{34}$$

Next, we replace $\frac{de^*_{ns}(s)}{ds}$ in $\frac{dw_s(s)}{ds}$ which yields,

$$\frac{dw_s(s)}{ds} = -\left[D'(E^*(s)) - \frac{C'_s\left(\psi^*_s\right)}{n}\right]\left[e^*_s(s) - e^*_{ns}(s)\right]\left[\frac{\Omega_s X_{ns}}{\Omega_s X_{ns} - (n-s)\Phi_{ns}X_s}\right]. \tag{35}$$

Since $D'(E^*(s)) - \frac{C'_s(\psi^*_s)}{n} > 0$ and $\frac{\Omega_s X_{ns}}{\Omega_s X_{ns} - (n-s)\Phi_{ns}X_s} > 0$, for all $s$, the sign of $\frac{dw_s(z)}{dz}$ depends only on $e^*_s(s) - e^*_{ns}(s)$.

Following the same steps as above, it can be shown that,

$$\frac{dw_{ns}(s)}{ds} = -\left[D'(E^*(s)) - \frac{C'_{ns}\left(\psi^*_s\right)}{n}\right]\left[e^*_s(s) - e^*_{ns}(s)\right]\left[\frac{\Omega_{ns} X_s}{\Omega_{ns} X_s - (n-1)\Phi_s X_{ns}}\right]. \tag{36}$$

Since $D'(E^*(s)) - \frac{C'_{ns}(\psi^*_s)}{n} > 0$ and $\frac{\Omega_{ns} X_s}{\Omega_{ns} X_s - (n-1)\Phi_s X_{ns}} > 0$, for all $s$, the sign of $\frac{dw_{ns}(z)}{dz}$ depends only on $e^*_s(s) - e^*_{ns}(s)$.

We know that, assuming completely identical countries, at $s^{\min} = 1$ signatories and nonsignatories choose the same level of emissions, $e^*_s(1) = e^*_{ns}(1) = e_N$. For $s \geqslant 2$, we know that signatories choose a lower level of emissions, $e^*_s(s) < e^*_{ns}(s)$ and the slope of both signatories and nonsignatories indirect welfare function with respect to $s$ is positive. For $s \geqslant 2$, even when signatories and nonsignatories are equaly responsive to the global norm, $\frac{dw_{ns}(s)}{ds} > \frac{dw_s(s)}{ds}$ due to the fact that $e^*_s(s) < e^*_{ns}(s)$. Thus, we can conclude that $w_s(s)$ is equal to $w_{ns}(s)$ at $s^{\min} = 1$ and for $s \geqslant 2$ both functions are increasing in $s$, with $w_{ns}(s)$ increasing faster than $w_s(s)$. ∎