International Environmental Agreements and the Paradox of Cooperation: Revisiting and Generalizing Some Previous Results

Michael Finus — *Department of Economics, Karl-Franzens-Universität Graz, Austria*
e-mail: michael.finus@uni-graz.at and Department of Economics, University of Bath, UK

Francesco Furini — *Department of Socioeconomics, Universität Hamburg, Germany*
e-mail: francesco.furini@uni-hamburg.de

Anna Viktoria Rohrer — *Department of Economics, Karl-Franzens-Universität Graz, Austria*
e-mail: anna.rohrer@uni-graz.at

Abstract

In his seminal paper Barrett (1994) argued that international environmental agreements (IEAs) are typical not successful, which he coined "the paradox of cooperation". Either self-enforcing IEAs are small and, hence, cannot achieve much or, if they are large, then the gains from cooperation are small. This message has been reiterated by several subsequent papers by and large. However, the determination of stable agreements and their evaluation have been predominantly derived for specific payoff functions and many conclusions are based on simulations. In this paper, we provide analytically solutions for the size of stable agreements, the paradox of cooperation and the underlying forces. Many of our results are a generalization of papers by Diamantoudi and Sartzetakis (2006), Rubio and Ulph (2006) and the recent paper by McGinty (2020).

Keywords: international environmental agreements, stability, paradox of cooperation

JEL-Classification: C72, D62, H41, Q50

1. Introduction

In his seminal paper Barrett (1994) argued that international environmental agreements (IEAs) are typical not successful, which he coined "the paradox of cooperation". Either self-enforcing IEAs are small and, hence, cannot achieve much or, if they are large, then the gains from cooperation are small. This message has been reiterated by several subsequent papers by and large.[1] However, the determination of stable agreements and their evaluation have been predominantly derived for specific payoff functions and many conclusions are based on simulations. In this paper, we provide analytically solutions for the size of stable agreements, the paradox of cooperation and the underlying forces. Many of our results are a generalization of later papers by Diamantoudi and Sartzetakis (2006), Rubio and Ulph (2006) and the recent paper by McGinty (2020).

Including Barrett (1994), all of these papers assume symmetric payoff functions for all countries and employ the workhorse model of IEAs which is the two-stage cartel formation game. In the first stage, countries decide about their membership. A coalition is called stable if those countries which have joined the coalition, called signatories, do not want to leave the agreement (internal stability) and those countries which have decided not to join the agreement, called non-signatories, do not want to join the agreement (external stability).[2] In the second stage, signatories choose their economic strategies (abatement or emissions) by maximizing the aggregate welfare of their members whereas non-signatories maximize their own welfare. Under the Nash-Cournot assumption, all countries choose their strategies simultaneously; under

[1] For a collection of some of the most influential papers and an overview article of those models, see Finus and Caparros (2015). Other overview articles include for instance Hovi et al. (2015) and Marrouch and Chauduri (2015).

[2] The concept has been borrowed from industrial economics (e.g., d'Aspremont et al. 1983). An alternative terminology of the cartel formation game is open membership single coalition game and internal and external stability is a Nash equilibrium in membership strategies (Yi 1997).

the Stackelberg assumption, signatories act as Stackelberg leaders and non-signatories as Stackelberg followers.

For most specific payoff functions, stable coalitions are small (compared to the total number of countries) under the Nash-Cournot assumption.[3] Hence, the pessimistic conclusion about the paradox of cooperation is obvious. However, the explanatory power of this model version is limited, as IEAs with large participation cannot be explained. In order to generate different results, some scholars have considered the Stackelberg assumption, which may lead to larger stable coalitions, including the grand coalition, depending on the benefit-cost structure of abatement.[4] All papers cited above in the text, including our paper, pursue this route.

Barrett (1994) central payoff function assumes quadratic benefits from global abatement and quadratic cost from individual abatement. Stable coalitions as well as the paradox of cooperation are illustrated with simulations. McGinty (2020) employs exactly the same payoff function. He introduces two effects, the externality and timing effect in order to provide a hint about the size of stable coalitions, which we denote by p^*. McGinty argues that both effects offset each other at a coalition of size $\hat{p}$. From his simulations he concludes that p^* is larger than $\hat{p}+1$ but strictly smaller than $\hat{p}+2$ and he confirms the paradox of cooperation.

For a general payoff function, we are able to characterize the externality and timing effect with reference to $\hat{p}$ and how this relates to p^*. We also provide a good approximation of the paradox

[3] An exception is Karp and Simon (2013), who develop a non-parametric model and consider non-standard abatement cost functions, like for instance concave marginal abatement cost functions or piecewise defined cost functions.

[4] Another possibility to generate different results is to stick to the Nash-Cournot assumption but to modify other assumptions by considering for instance modest emission reduction targets (Finus and Maus 2008), asymmetric countries (Finus and McGinty 2019, Fuentes-Albero and Rubio 2010 and Pavlova and de Zeeuw 2013) and additional strategies like R&D (e.g., Barrett 2006, El-Sayed and Rubio 2014, Hoel and de Zeeuw 2010 and Rubio 2017) or adaptation (e.g., Bayramoglu et al. 2018 and Rubio 2018).

of cooperation. For his specific payoff function, we analytically determine p^* and measure the paradox of cooperation and relate it to the benefit and cost parameter of the model.

Diamantoudi and Sartzetakis (2006) as well as Rubio and Ulph (2006) transform Barrett's payoff function in abatement space to the dual problem in emission space. They show that complications arise if one imposes the constraint that emission have to be non-negative. Diamantoudi and Sartzetakis (2006) impose parameter constraints in order to ensure only interior solutions. This implies that the model no longer predicts $p^* \in [2,n]$, with n the total number of countries, but only $p^* \in [2,4]$.[5] In contrast, Rubio and Ulph (2006) work with Kuhn-Tucker conditions in order to ensure non-negative emissions. They confirm $p^* \in [2,n]$ and the paradox of cooperation via simulations; they are able to analytical characterize parameter ranges for some values of p^*, though not for the entire parameter space.

In contrast, we work with a model in abatement space for which non-negativity conditions cause less of a problem for analytical solutions. As pointed out above, we provide a full and exact analytical characterization of p^* as well as for the paradox of cooperation for the entire parameter space of the model. Even for a general payoff function, we are able to provide a good approximation of those features. Finally, we provide a general proof that p^* is at least as large under the Stackelberg than under the Nash-Cournot assumption, a conclusion, which, to the best of our knowledge, has only been derived from simulations until now. This relation also motivates why we mainly focus on the Stackelberg assumption in this paper.

---

[5] Diamantoudi and Sartzetakis (2006) already determine $\hat{p}$, how this relates to the payoff of signatories and non-signatories and that $\hat{p}+1$ is internally stable for their specific payoff function provided non-negative emissions are ignored, something which seems to have been unnoticed by McGinty (2020). We are able to establish all these features for a general payoff function.

2. The Model

There are n symmetric countries $i = 1,2,...,n$, with N being the set of all countries. In the first stage of the game, countries decide whether to join coalition $P \subseteq N$ and become signatories (S) or to remain outside as non-signatories (NS). The size of coalition P is denoted by p. In the second stage of the game, countries select their mitigation levels. Players in the coalition will maximize their aggregate welfare, while players outside of the coalition will maximize their individual welfare. The individual welfare function is given by

$$W_i(M, m_i) = B_i(M) - C_i(m_i). \tag{1}$$

Benefits B_i arise from total mitigation $M = \sum_{i=1}^{n} m_i$, while costs C_i depend on individual mitigation m_i. (We use the terms mitigation and abatement as synonyms.)

All components of the welfare function are assumed to be continuous, including its first and second derivatives. The benefit function is assumed to be increasing in total mitigation at a decreasing rate $(B_M > 0$ and $B_{MM} < 0)$, while costs are assumed to be a strictly convex function of individual mitigation $(C_m > 0$ and $C_{mm} > 0)$ where subscripts refer to derivatives, e.g., $B_M = \partial B_i / \partial M$ and $B_{MM} = \partial^2 B_i / \partial M^2$. Subsequently, we will sometimes drop the arguments in some functions for notational simplicity. In order to guarantee interior solutions, one may impose for instance the condition $\lim_{M \to 0} B_M > \lim_{m \to 0} C_m > 0$. Moreover, in order to avoid signing third derivatives and to simplify the mathematics, we assume that second derivatives are constant. The game is solved by backward induction.

In the second stage, there are two possible versions of the game, depending on the sequence of decisions. In the Nash-Cournot (NC-) scenario, signatories and non-signatories choose their

mitigation strategies simultaneously. In the Stackelberg (ST-) scenario, mitigation levels are first set by signatories (taking into account the best-response of non-signatories) and then are chosen by non-signatories. In case no coalition has formed in the first stage of the game, i.e., $p=1$, it is assumed that a randomly selected country will behave as a leader while all other $n-1$ countries will be the followers.[6] For the two scenarios the first order conditions in an interior solution are displayed in Table 1. We note that in both scenarios the second order conditions automatically hold due to the assumption about the properties of the benefit and cost functions and a unique second stage equilibrium exists (see Appendix A.1). Thus, instead of writing $W_i\left(M^*(p), m_i^*(p)\right)$, we can simply write $W_S^*\left(p\right)$ and $W_{NS}^*\left(p\right)$ where the asterisks indicate equilibrium values for a given coalition size p, $1 \le p \le n$.

Table 1: First Order Conditions in the NC- and ST-scenario

	NC-scenario	ST-scenario
Signatories	$p \cdot B_M\left(M^{NC*}\right) = C_m\left(m_S^{NC*}\right)$ (2)	$p \cdot \left(1+R_{NS}^{'}\right)\left[B_M\left(M^{ST*}\right)\right] = C_m\left(m_S^{ST*}\right)$ (4)
Non-signatories	$B_M\left(M^{NC*}\right) = C_m\left(m_{NS}^{NC*}\right)$ (3)	$B_M\left(M^{ST*}\right) = C_m\left(m_{NS}^{ST*}\right)$ (5)

The first order conditions implicitly define individual best-response functions of signatories, $m_{i\in P} = r_S(M_{-i})$, and non-signatories, $m_{j\notin P} = r_{NS}(M_{-j})$, with M_{-i} and M_{-j} being the aggregate mitigation level of all countries except country $i \in P$ and $j \notin P$, respectively. As countries are symmetric, we can define the aggregate best-response function of signatories $M_S = R_S(M_{NS})$ and non-signatories $M_{NS} = R_{NS}(M_S)$, with M_{NS} being the aggregate mitigation level of all non-

[6] We make this assumption to be in line with McGinty (2020), even though the alternative assumption, namely that only above $p \ge 2$ signatories assume Stackelberg leadership and for $p=1$ the ST- and NC-scenario are identical, would almost always lead to exactly the same results.

signatories and M_S the aggregate mitigation level of all signatories. Accordingly, the slopes of those reaction functions are given by

$$r_S^{'}\left(M_{-i\in P}\right)=\frac{p\cdot B_{MM}}{C_{mm}\left(m_S\right)-p\cdot B_{MM}}, \qquad (6) \qquad R_S^{'}\left(M_{NS}\right)=\frac{p^2\cdot B_{MM}}{C_{mm}\left(m_S\right)-p^2\cdot B_{MM}}, \qquad (7)$$

$$r_{NS}^{'}\left(M_{-j\notin P}\right)=\frac{B_{MM}}{C_{mm}\left(m_{NS}\right)-B_{MM}} \qquad (8) \text{ and } \qquad R_{NS}^{'}\left(M_S\right)=\frac{\left(n-p\right)\cdot B_{MM}}{C_{mm}\left(m_{NS}\right)-\left(n-p\right)\cdot B_{MM}} \ . \qquad (9)$$

We note that all denominators are positive due to the second order conditions (see Appendix A.1). Because $B_{MM}<0$, all reaction functions are negatively sloped. That is, mitigation levels are strategic substitutes and the slope can be interpreted as a measure of the leakage effect where all slopes lie in the interval $[-1,0]$. The absolute values of these slopes increase in the absolute value of B_{MM} and decrease in the value C_{mm}, with the slopes approaching -1 if B_{MM}/C_{mm} goes to infinity and 0 if B_{MM}/C_{mm} goes to zero.

In the first stage of the game, a coalition of size p is stable, denoted by p^*, if it is simultaneously internally

$$W_S^*\left(p^*\right)\geq W_{NS}^*\left(p^*-1\right) \qquad (10)$$

and externally stable

$$W_{NS}^*\left(p^*\right)\geq W_S^*\left(p^*+1\right). \qquad (11)$$

Alternatively, we can define the stability function $\Omega(p):=W_S^*\left(p\right)-W_{NS}^*\left(p-1\right)$. Then, a coalition of size p^* is stable if $\Omega(p^*)\geq 0$ and $\Omega(p^*+1)\leq 0$ hold simultaneously. We can also conclude that if a coalition of size p is strictly internally stable (i.e., $\Omega(p)>0$), then the coalition of size $p-1$ is externally unstable (i.e., $\Omega(p-1)>0$).

3. General Results

For the following analysis, it is useful to note that from the first order conditions in the NC-scenario (see eqs. (2) and (3) in Table 1), and those in the ST-scenario (see eqs. (4) and (5) in Table 1), we have:

$$C_m\left(m_S^{NC*}\right) = p \cdot C_m\left(m_{NS}^{NC*}\right) \text{ and } C_m\left(m_S^{ST*}\right) = p \cdot \left(1 + R'_{NS}\right) \cdot C_m\left(m_{NS}^{ST*}\right) \tag{12}$$

where we may recall that R'_{NS} denotes the slope of the aggregate reaction function of non-signatories (see eq. (9)). Given that mitigation costs are strictly convex, $m_S^{NC*}(p) > m_{NS}^{NC*}(p)$ follows in the NC-scenario immediately for every p, $1 < p < n$, as stated in Proposition 1 below. This is different in the ST-scenario. We first note that $p \cdot \left(1 + R'_{NS}\right)$ is a short-hand notation for $p \cdot \left(1 + R'_{NS}(p)\right)$, i.e., the slope of the reaction function is also a function of p. Moreover, we note that $p \cdot \left(1 + R'_{NS}\right) < 1$ for $p = 1$ and that $p \cdot \left(1 + R'_{NS}\right)$ increases in p, as we show in Appendix A.2. Therefore, there exists a $\hat{p}$ for which $\hat{p} \cdot \left(1 + R'_{NS}\right) = 1$ holds, with $1 < \hat{p} < n$. Hence, $m_S^{ST*}(p) < m_{NS}^{ST*}(p)$ if $p < \hat{p}$, and if $\hat{p} < n$, then $m_S^{ST*}(p) \geq m_{NS}^{ST*}(p)$ if $p \geq \hat{p}$. Since all countries have the same benefits, which only depend on total mitigation, differences in welfare levels between signatories and non-signatories stem from different mitigation levels and, hence, show up in different mitigation costs. This covers part i) and ii) of Proposition 1.

Proposition 1

Consider a generic coalition of size p. *The following relations hold:*

i) Nash-Cournot Scenario

$m_S^{NC*}(p) > m_{NS}^{NC*}(p)$ *and* $W_S^{NC*}(p) < W_{NS}^{NC*}(p)$ *for any* p, $1 < p < n$.

ii) Stackelberg Scenario

$m_S^{ST*}(p) < (\geq) m_{NS}^{ST*}(p)$ *and* $W_S^{ST*}(p) > (\leq) W_{NS}^{ST*}$ *if* $p < (\geq) \hat{p}$, *for any* p, $1 \leq p < n$.

iii) Comparison across Scenarios

$m_S^{NC*}(p) > m_S^{ST*}(p)$, $m_{NS}^{NC*}(p) < m_{NS}^{ST*}(p)$ *and* $M^{NC*}(p) > M^{ST*}(p)$;

$W_S^{NC*}(p) \leq W_S^{ST}(p)$ *and* $W_{NS}^{NC*}(p) > W_{NS}^{ST*}(p)$ *for any* p, $1 \leq p < n$.

Proof: i) and ii) follow from the discussion above; iii) is proved in Appendix A.3. **Q.E.D.**

From part i) in Proposition 1 it is evident that a signatory's mitigation level is always larger than a non-signatory's mitigation level for any non-trivial coalition in the NC-scenario. Consequently, a signatory's welfare level will always be smaller than a non-signatory's welfare level, which already hints at why stable coalitions tend to be small. From part ii) in Proposition 1 it emerges that these relations only hold for larger coalitions above $\hat{p}$ in the ST-scenario, but this is reversed for smaller coalitions below $\hat{p}$, providing already some intuition why stable coalitions tend to be larger in the ST- than NC-scenario. This is further substantiated in part iii) in Proposition 1. The Stackelberg leaders use their strategic advantage: they reduce their mitigation levels compared to the NC-scenario, knowing that followers will compensate for this to some extent. As compensation is incomplete, due to the fact that the slopes of the reaction functions are strictly larger than -1, total mitigation levels will be smaller in the ST- than in

the NC-scenario. This strategic shift also shows up in the relation of welfare levels between both scenarios. The immediate and central implication is summarized in Corollary 1.

Corollary 1

- *Stable coalitions in the ST-scenario are weakly larger than in the NC-scenario: That is,* $p^{ST*} \geq p^{NC*}$.

Proof: From Proposition 1, part iii), we have $W_S^{NC*}(p) \leq W_S^{ST}(p)$ and $W_{NS}^{NC*}(p) > W_{NS}^{ST*}(p)$ for any p, $1 \leq p < n$. Thus, also $W_{NS}^{NC*}(p-1) > W_{NS}^{ST*}(p-1)$ and, consequently, $\Omega^{ST}(p) \geq \Omega^{NC}(p)$. Let $p = \vec{p}$ be the largest coalition which is internally and externally stable in the NC-scenario. Either $p = \vec{p}$ is also externally stable in the ST-scenario or if not, then there will be a large coalition which is internally and externally stable, knowing that the grand coalition is externally stable for sure. **Q.E.D.**

In order to provide an intuition for Corollary 1, McGinty (2020) suggests to consider two effects. He calls the first effect the *externality effect* which he defines as follows: $EE^M = m_S^{NC*}(p) - m_{NS}^{NC*}(p-1)$. This effect measures to which extent signatories mitigate more than non-signatories in the NC-scenario and which causes free-riding. He calls the second effect the *timing effect*, which reduces the incentive to free-ride in the ST- compared to the NC-scenario. He captures the timing effect in two dimensions: $TE_1^M = m^{NE*} - m_S^{ST*}(1)$ and $TE_2^M = m_{NS}^{ST*}(1) - m^{NE*}$ with the superscript NE referring to mitigation level in the Nash equilibrium, which, in the NC-scenario, is equal to mitigation levels for $p = 1$, i.e., $m_S^{NC*}(1) = m_{NS}^{NC*}(1) = m^{NE*}$. Hence, TE_1^M measures the extent by which the single Stackelberg leader reduces its mitigation level compared to the Nash equilibrium and TE_2^M measures the extent by which each of the $n-1$ followers increase their mitigation levels compared to the

Nash equilibrium; in both cases $p=1$ is considered. McGinty (2020) argues that the two effects offset each other at $\widehat{p}$ in the ST-scenario.

We agree with McGinty (2020) that these two effects are useful in explaining the location of $\hat{p}$, but we disagree with the definitions of those effects. First, measuring the externality effect at different values of p for signatories and non-signatories appears somehow inconsistent. Second, measuring the timing effect only for $p=1$ does not appear convincing to explain the location of $\widehat{p}$ for which typically $\widehat{p}>1$ holds. Third, arguing that the two effects set off each other at $\widehat{p}$ suggests that aggregating the two effects in one way or the other should give us exactly $\widehat{p}$. Fourth, the effects should not only be measured in mitigation space but also in welfare terms. Therefore, we suggest the following definitions, which we use henceforth.

Externality Effect: $EE^M = m_S^{NC^*}(p) - m_{NS}^{NC^*}(p)$ and $EE^W = W_S^{NC^*}(p) - W_{NS}^{NC^*}(p)$.

Timing Effect: $TE^M = m_S^{NC^*}(p) - m_S^{ST^*}(p) + m_{NS}^{ST^*}(p) - m_{NS}^{NC^*}(p)$ and

$$TE^W = W_S^{NC^*}(p) - W_S^{ST^*}(p) + W_{NS}^{ST^*}(p) - W_{NS}^{NC^*}(p)$$

Total Effect: $ToE^M = EE^M - TE^M = m_S^{ST^*}(p) - m_{NS}^{ST^*}(p)$ and

$$ToE^W = EE^W - TE^W = W_S^{ST^*}(p) - W_{NS}^{ST^*}(p)$$

From the externality effect, responsible for free-riding, implying small stable coalitions in the NC-scenario, the timing effect, reducing the free-rider incentive, is deducted in order to obtain the total effect. Hence, we can state the following.

Corollary 2

Consider any coalition of size p, $1 \le p < n$. *Let* $p = \hat{p}$ *be such that* $\hat{p} \cdot \left(1 + R'_{NS}(\hat{p})\right) = 1$, $p \cdot \left(1 + R'_{NS}(p)\right) < 1$ *if* $p < \hat{p}$ *and* $p \cdot \left(1 + R'_{NS}(p)\right) > 1$ *if* $p > \hat{p}$, *with* $1 < \hat{p} \le n$.

a. *The externality effect in mitigation space is positive, i.e.,* $EE^M > 0$ *and negative in welfare space, i.e.,* $EE^W < 0$.
b. *The timing effect in mitigation space is positive, i.e.,* $TE^M > 0$ *and negative in welfare space, i.e.,* $TE^W < 0$.
c. *The total effect in mitigation space is positive (negative) if* $p \le \hat{p}$ $(p > \hat{p})$, *i.e.,* $ToE^M \ge (<) 0$ *if* $p \le \hat{p}$ $(p > \hat{p})$ *and positive (negative) in welfare space if* $p \le \hat{p}$ $(p > \hat{p})$, *i.e.,* $ToE^W \ge (<) 0$ *if* $p \le \hat{p}$ $(p > \hat{p})$. *That is, the two effects offset each other at* $\hat{p}$.

Proof: Part a. and b. follow directly from Proposition 1; part c. follows immediately from the first order conditions (4) and (5) in Table 1 and the properties of $p \cdot \left(1 + R'_{NS}(p)\right)$ as discussed above and established in Appendix A.2. **Q.E.D.**

In a next step, we have closer look how the coalition size p and in particular the benchmark value $p = \hat{p}$ relate to mitigation and welfare levels in the ST-scenario. We focus on this scenario, given that we have established $p^{ST*} \ge p^{NC*}$ in Corollary 1.

Proposition 2

Consider any coalition of size p, $1 < p < n$, *recall the definition of* $\hat{p}$. *Let NE denote Nash equilibrium mitigation and welfare levels.*

a. $m_S^{ST*}\left(\hat{p}\right) = m_{NS}^{ST*}\left(\hat{p}\right) = m^{NE*}$ *and* $M^{ST*}\left(\hat{p}\right) = M^{NE*}$. *For* $p < (>)\,\hat{p}$, $m_S^{ST*}(p) < (>)\, m^{NE*}$, $m_{NS}^{ST*}(p) > (<)\, m^{NE*}$ *and* $M^{ST*}(p) < (>)\, M^{NE*}$.

b. $W_S^{ST*}\left(\hat{p}\right) = W_{NS}^{ST*}\left(\hat{p}\right) = W_i^{NE*}$ *and* $W^{ST*}\left(\hat{p}\right) = W^{NE*}$. *For* $p < (>)\,\hat{p}$, $W_{NS}^{ST*}(p) < (>)\, W_i^{NE*}$ *and* $W^{ST*}(p) < (>)\, W^{NE*}$. *Finally,* $W_S^{ST*}(p) \geq W_i^{NE*}$ *for every* $p \in [1, n]$.

c. *For* $1 \leq p < \hat{p}$, $m_{NS}^{ST*}(p)$ *decreases,* $M^{ST*}(p)$ *increases and* $m_S^{ST*}(p)$ *increases in* p *in some segment of this range. For* $\hat{p} \leq p < n$, $m_{NS}^{ST*}(p)$ *continuously decreases and* $M^{ST*}(p)$ *continuously increases in* p. $m_S^{ST*}(p)$ *increases in* p *in some segment of this range.*

d. *For* $1 \leq p < \hat{p}$, $W_{NS}^{ST*}(p)$ *and* $W^{ST*}(p)$ *increase and* $W_S^{ST*}(p)$ *decreases in* p *in some segment of this range. For* $\hat{p} \leq p < n$, $W_{NS}^{ST*}(p)$, $W_S^{ST*}(p)$ *and* $W^{ST*}(p)$ *continuously increase in* p.

Proof: See Appendix A.4. **Q.E.D.**

Proposition 3 is illustrated in Figure 1 with two representative cases. Plots 1.a to 1.d imply a relatively low $\hat{p}$ and plots 1.e to 1.h imply a relatively high $\hat{p}$.

In terms of individual mitigation and welfare levels (plots 1.a, 1.c, 1.e and 1.g), it is clear that the position of $\hat{p}$ determines the magnitude of the strategic advantage of signatories over non-signatories in the ST-scenario. From $p = 1$ up to shortly before $p = \hat{p}$, the timing effect dominates the externality effect; signatories will mitigate less than non-signatories and less than in the Nash equilibrium. Moreover, they receive a higher payoff compared to non-signatories. Non-signatories mitigate more than in the Nash equilibrium but receive a lower payoff than in the Nash equilibrium. For $p > \hat{p}$, signatories and non-signatories receive a higher welfare level

than in the Nash equilibrium and both groups' welfare increases in the coalition size p. However, now the externality effect dominates the timing effect, implying that signatories have higher mitigation but lower welfare than non-signatories.

Given that signatories' welfare is larger and non-signatories' welfare lower than in the Nash equilibrium for any $p < \widehat{p}$, welfare (and mitigation) levels are equal to Nash equilibrium levels for $p = \widehat{p}$ and signatories' payoffs increase in p for any $p > \widehat{p}$, we know for sure that all coalitions up to $p = \widehat{p} + 1$ are internally stable (because in this range $W_S(p) \geq W_S(p-1) \geq W_{NS}(p-1)$). Acknowledging the fact that the size of stable coalitions can take on only integer values, we define $\lfloor p \rfloor = \max\{z \in \mathbb{Z} | z \leq p\}$.

Corollary 3

Consider the ST-scenario. Let $\widehat{p} + 1 < n$. *Every coalition of size* $p \in \left[2, \left\lfloor \widehat{p} + 1 \right\rfloor\right]$ *is internally stable. Hence, every coalition of size* $p < \left\lfloor \widehat{p} + 1 \right\rfloor$ *is externally unstable. Therefore* $p^* \geq \left\lfloor \widehat{p} + 1 \right\rfloor$. *If* $\widehat{p} + 1 \geq n$, *then,* $p^* = n$.

At this level of generality, nothing can be concluded whether coalitions for which $p \geq \left\lfloor \widehat{p} + 2 \right\rfloor$ holds will be stable if $\widehat{p} + 1 < n$. For the specific welfare function (13), which we consider in section 4, it turns out that coalitions for which $p > \left\lfloor \widehat{p} + 2 \right\rfloor$ holds are not stable. Nevertheless, Proposition 2 and Corollary 2 already provide a very good intuition about the paradox of cooperation.

First note that because $\widehat{p} \cdot \left(1 + R'_{NS}(\widehat{p})\right) = 1$, $\widehat{p}$ increases in the absolute value of the slope of the reaction function R'_{NS}. That is, the steeper the negatively sloped reaction function is, the larger

will be $\hat{p}$. Hence, because $p^* \geq \left\lfloor \hat{p}+1 \right\rfloor$, it appears that the steeper reaction functions are, the larger will be stable agreements. If R'_{NS} is sufficiently steep, the grand coalition will be stable.

However, global welfare and mitigation are below those in the Nash equilibrium for $p < \hat{p}$, and are only large (and increase in p) above the Nash equilibrium for $p > \hat{p}$. Consequently, a large $\hat{p}$ close to n does not allow that a stable coalition equal or larger than $\hat{p}+1$ improves much over the Nash equilibrium. Indeed, if $\hat{p}$ is close to n, not much room is left for additional countries to join the coalition in order to improve over the Nash equilibrium.

This is also evident from Figure 1, when comparing plots 1.b and 1.f in terms of global mitigation and plots 1.d and 1.h in terms of global welfare. In the top panels, $\hat{p}$ is relatively small but the potential gains from cooperation would be large; in the lower panels, this is just reversed.

4. The Paradox of Cooperation

4.1 Preliminaries

In order to obtain some more concrete results regarding the paradox of cooperation, we consider the welfare function in Barrett (1994), which has also been considered by McGinty (2020) and many others.

$$w_i = \frac{b}{n}\left(a \cdot M - \frac{M^2}{2}\right) - \frac{c}{2} \cdot m_i^2 \tag{13}$$

a, b and c are strictly positive parameters, n denotes the total number of countries, M stands for global mitigation and m_i for individual mitigation. All detailed calculations are provided in in Online Appendix 1. In order for the benefit function to be in line with our general assumptions

in section 2, we need to require $M < a$ for $B_M > 0$. It turns out that this upper bound never becomes binding for equilibrium values for every p, $1 \le p \le n$.

For the slopes of the reaction functions, as derived in section 2 for the general welfare function (1) (see eqs. (6) to (9)), using $\gamma = \frac{c}{b}$ as in Barrett (1994), we find for payoff function (13):

$$r_S'\left(M_{-i\in P}\right) = -\frac{p}{\gamma \cdot n + p}, \qquad R_S'\left(M_{NS}\right) = -\frac{p^2}{\gamma \cdot n + p^2},$$

$$r_{NS}'\left(M_{-j\notin P}\right) = -\frac{1}{\gamma \cdot n + 1} \text{ and} \qquad R_{NS}'\left(M_S\right) = -\frac{(n-p)}{\gamma \cdot n + (n-p)}.$$

It is straightforward to see that all reaction functions become steeper (flatter) as γ decreases (increases). For γ going to infinity, the slopes go to 0. For γ going to 0, the slopes approach the value of -1.

The critical coalition size $\hat{p}$, as defined in section 3, at which the timing and externality effects offset each other, is given by:

$$\hat{p} = \frac{n \cdot (1+\gamma)}{n \cdot \gamma + 1}. \tag{14}$$

The value of $\hat{p}$ is the same as McGinty (2020). We note that $\hat{p}$ decreases in γ. Hence, $\hat{p}$ moves with the absolute value of the slope of the reaction functions.

In order to analyze the paradox of cooperation, we propose relative measures to evaluate the potential gains from cooperation in ecological and welfare terms. We believe that relative measures are more sensible than absolute measures, which is in particular true for welfare. We call our indexes Importance of Cooperation Indexes (ICI), abbreviated M-ICI in terms of *M*itigation and W-ICI in terms of *W*elfare. They measure the difference between the social optimum (SO) and the non-cooperative Nash equilibrium (NE) in relation to the levels in the

Nash equilibrium. (Note that the social optimum is identical for both scenarios if a coalition of size $p = n$ forms, the grand coalition.

$$W - ICI = \frac{W^{SO*} - W^{NE*}}{W^{NE*}} \qquad M - ICI = \frac{M^{SO*} - M^{NE*}}{M^{NE*}}$$

For the welfare function (13), we obtain:

$$W - ICI = \frac{(n-1)^2 \cdot c^2}{(n \cdot b + c) \cdot (n \cdot b + 2n \cdot c - c)} \quad (15) \text{ and } M - ICI = \frac{(n-1) \cdot c}{n \cdot b + c}. \quad (16)$$

It is straightforward to confirm that both indexes increase in parameter c and decreases in the parameter b, just the opposite than what has been observed for the absolute slope of the reaction functions.

4.2 Results

For payoff function (13), we obtain the following result.

Proposition 3

In the ST-scenario, the unique equilibrium coalition size p^* *is given by* $p^* \in [2, n]$. *Let* $\hat{p} + 1 < n$, *i.e.,* $\gamma > \frac{1}{n[n-2]}$. *If* $\hat{p}$, *as given in (14), is not an integer value, the unique stable coalition size is one of two integer values,* $p^* \in \{\lfloor \hat{p} + 1 \rfloor, \lfloor \hat{p} + 2 \rfloor\}$, *whereas if* $\hat{p}$ *is an integer value, then the unique stable coalition size is* $p^* = \hat{p} + 1$. *If* $\hat{p} + 1 \geq n$, *i.e.,* $\gamma \leq \frac{1}{n[n-2]}$, *then* $p^* = n$.[7]

[7] In the NC-scenario, we would have $p^* \in [1, 2]$, which can be proved along the lines as for instance developed in Appendix A.3.3 in Bayramoglu et al. (2018), who consider almost the same welfare function as in (13).

Proof: For the ST-scenario, see Appendix A.4 for details where we show (treating p as a continuous variable) that $p = \widehat{p} + 2$, and any larger coalition, is not internally stable (though any coalition from $p = 2$ up to $p = \widehat{p} + 1$ is internally stable as we know from Corollary 2 already) and the range of γ follows directly from (14) and the conditions $\widehat{p} + 1 < n$ and $\widehat{p} + 1 \geq n$, respectively. Finally, results are obtained by acknowledging the fact that p^* must be an integer value. **Q.E.D.**

It does not come as a surprise that we can derive a sharper characterization of stable coalitions for the specific welfare function (13) than for the general welfare function (1) in Corollary 2. It is interesting that our previous intuition for the general welfare function (1) about the relation between the size of stable coalitions and the paradox of cooperation is confirmed for the specific payoff function (13).

Corollary 4

The unique equilibrium coalition size p^* increases in the absolute values of the slopes of the reaction functions which in turn increase in the benefit parameter b and decreases in the cost parameter c whereas the importance of cooperation in welfare and mitigation terms (see eqs. (15) and (16)) decrease in the benefit parameter b and increases in the cost parameter c.

Proof: Follows from the discussion above and Proposition 3. **Q.E.D.**

Thus, $\widehat{p}$ and, hence, p^*, increases in the absolute value of the slopes of the reaction functions (which increase in parameter b and decrease in parameter c). In terms of the gains from cooperation, measured by our indexes of the importance of cooperation, just the reverse is true. This constitutes the paradox of cooperation.

5 Conclusion

The paradox of cooperation is a well-established result from the literature on international environmental agreements games: either self-enforcing environmental agreements comprise only few countries or, if large participation is achieved, cooperation is not able to substantially improve environmental conditions and welfare. This conclusion, since its first formulation by Barrett (1994), has been reiterated in the literature by many scholars. However, results have been based on specific payoff functions, and to a large extent on simulations (e.g., Diamantoudi and Sartzetakis 2006, McGinty 2020 and Rubio and Ulph 2006). We derive many results for general payoff functions, and some further analytical results for a specific payoff function, frequently employed in this literature.

In this paper, we first proved generally that stable coalitions are larger under the Stackelberg leadership than under the Nash-Cournot assumption. Given that stable coalitions under the Nash-Cournot assumption always tend to be small, we focused on the Stackelberg assumption in the subsequent analysis. We characterized the range of stable coalitions and provided a rationale for the paradox of cooperation. The steeper reaction functions are, the larger is the strategic advantage of signatories over non-signatories and, consequently, the larger are stable coalitions, but the smaller are the gains from cooperation. We argued that this simple relation summarizes the paradox of cooperation in a nutshell. We generalized the timing and externality effect, as suggested by McGinty (2020), in order to locate stable coalitions. The final analysis of the specific payoff function allowed us to relate all previous results to the ratio of the benefit and cost parameter in this public good coalition formation model.

For further research, we suggest that more efforts are undertaken to generalize previous results, in order to draw more robust conclusions.

References

Barrett, S., (1994), Self-Enforcing International Environmental Agreements. "Oxford Economic Papers", 46, pp.878–894.

Barrett, S. (2006), Climate Treaties and "Breakthrough" Technologies. "American Economic Review", vol. 96(2), pp. 22-25.

Bayramoglu, B., M. Finus and J.-F. Jacques (2018), Climate Agreements in Mitigation-Adaptation Game. "Journal of Public Economics", vol. 165, pp. 101-113.

Cornes, R.C. and R. Hartley (2007), Aggregative Public Good Games. "Journal of Public Economic Theory", vol. 9 (2), pp. 201–219.

D'Aspremont, C., A. Jacquemin, J. Gabszewicz and J. Weymark (1983), On the Stability of Collusive Price Leadership. "The Canadian Journal of Economics", vol. 16(1), pp. 17-25.

Diamantoudi, E. and E.S. Sartzetakis (2006), Stable International Environmental Agreements: An Analytical Approach. "Journal of Public Economic Theory", vol. 8(2), pp. 247-263.

El-Sayed, A. and S. Rubio (2014), Sharing R&D Investments in Cleaner Technologies to Mitigate Climate Change. "Resource and Energy Economics", vol. 38, pp. 168-180.

Finus, M. and A. Caparrós (2015), Game Theory and International Environmental Cooperation. The International Library of Critical Writings in Economics. Edward Elgar, Cheltenham, UK.

Finus, M. and S. Maus. (2008), Modesty May Pay! "Journal of Public Economic Theory", vol. 10, pp. 801–26.

Finus, M. and M. McGinty (2019), The Anti-Paradox of Cooperation: Diversity May Pay. "Journal of Economic Behavior & Organization", Vol. 157, pp. 541-559.

Fuentes-Albero, C. and S.J. Rubio (2010), Can International Environmental Cooperation be Bought? "European Journal of Operational Research", vol. 202, pp. 255–264

Hoel, M. and A. de Zeeuw (2010), Can a Focus on Breakthrough Technologies Improve the Performance of International Environmental Agreements? "Environmental and Resource Economics", vol. 47, pp. 395-406.

Hovi, J., H. Ward and F. Grundig (2015), Hope or Despair? Formal Models of Climate Cooperation. "Environmental and Resource Economics", vol. 62, pp. 665–688.

Karp, L. and L. Simon (2013), Participation Games and International Environmental Agreements: A Non-Parametric Model. "Journal of Environmental Economics and Management", vol. 65(2), pp. 326-344.

Marrouch, W. and A.R. Chaudhuri (2015), International Environmental Agreements: Doomed to Fail or Destined to Succeed? A Review of the Literature. "International Review of Environmental and Resource Economics", vol. 9, pp. 245–319.

McGinty, M. (2020), Leadership and Free-Riding: Decomposing and Explaining the Paradox of Cooperation in International Environmental Agreements. "Emvironmental and Resource Economics", vol. 77, pp. 449-474.

Pavlova, Y. and A. de Zeeuw (2013), Asymmetries in International Environmental Agreements. "Environment and Development Econonomics", vol. 18, pp. 51–68.

Rubio, S.J. and A. Ulph (2006), Self-enforcing International Environmental Agreements Revisited. "Oxford Economic Papers", vol. 58(2), pp. 233-263.

Rubio, S. (2017), Sharing R&D Investments in Breakthrough Technologies to Control Climate Change. "Oxford Economic Papers", vol. 69(2), pp. 496-521.

Rubio, S.J. (2018), Self-Enforcing International Environmental Agreements: Adaptation and Complementarity. Working Paper, 029.2018, Fondazione Eni Enrico Mattei.

Yi, S.S. (1997), Stable Coalition Structures with Externalities. "Games and Economic Behavior", vol. 20, pp. 201–237.

Appendix

A.1 Second order conditions and the existence of a unique second stage equilibrium

Differentiating the first order conditions in Table 1 with respect to individual mitigation, we obtain the following SOCs in the NC-scenario: $p^2 \cdot B_{MM}\left(M^{NC*}\right) - C_{mm}\left(m_S^{NC*}\right) < 0$ for signatories and $B_{MM}\left(M^{NC*}\right) - C_{mm}\left(m_{NS}^{NC*}\right) < 0$ for non-signatories. In the ST-scenario, SOCs for non-signatories are the same as in the NC-scenario (replacing superscripts NC by ST). For signatories, we obtain: $p^2 \cdot \left(1 + R_{NS}^{'}\right) \cdot B_M\left(M^{ST*}\right) - C_{mm}\left(m_S^{ST*}\right) < 0$ where for simplicity we assume constant second derivatives.

The existence of a unique vector of mitigation levels for every coalition of size p is proved by using the concept of replacement functions (see Cornes and Hartley 2007). Let $m_S = g_S(M)$ be the individual replacement function of a signatory and $m_{NS} = g_{NS}(M)$ be the replacement function of a non-signatory. The aggregate replacement function $G(M)$ is the summation of all replacement functions:

$$\sum_{i=1}^{n} m_i = p \cdot m_S + (n-p) \cdot m_{NS} = M = G(M) = \sum_{i=1}^{n} g_i(M) = p \cdot g_S(M) + (n-p) \cdot g_{NS}(M).$$

For the ST-scenario, totally differentiating the first order conditions of signatories and of non-signatories (eqs. (4) and (5), respectively), we derive the slope of an individual signatory's and non-signatory's replacement function:

$$g_S^{'}\left(M^{ST}\right) = \frac{p \cdot \left[B_{MM} \cdot \left(1 + R_{NS}^{'}\right)\right]}{C_{mm}\left(m_S^{ST}\right)} \qquad \text{and} \qquad g_{NS}^{'}\left(M^{ST}\right) = \frac{B_{MM}}{C_{mm}\left(m_{NS}^{ST}\right)}.$$

The slope of the aggregate replacement function is obtained by summing over all individual slopes:

$$G^{'}\left(M^{ST}\right) = B_{MM} \cdot \left[\frac{p^2 \cdot \left[\left(1 + R_{NS}^{'}\right)\right]}{C_{mm}\left(m_S^{ST}\right)} + \frac{(n-p)}{C_{mm}\left(m_{NS}^{ST}\right)}\right].$$

Both individual and aggregate replacement functions have a negative slope over the entire domain of the mitigation space as $B_{MM} < 0$. Hence, the aggregate replacement function will

intersect with the 45-degree line only once. For the NC-scenario, the same result is obtained by setting $R'_{NS} = 0$ and replacing the superscript ST by NC.

A.2 The Nature of $\hat{p}$

Consider $p \cdot \left(1 + R'_{NS}(p)\right)$. Since $R'_{NS}(p) < 0$, $p \cdot \left(1 + R'_{NS}(p)\right) < 1$ for $p = 1$ and hence $\hat{p} > 1$ with $\hat{p} \cdot \left(1 + R'_{NS}\left(\hat{p}\right)\right) = 1$. $p \cdot \left(1 + R'_{NS}(p)\right)$ increases in p. To see this, we differentiate R'_{NS} with respect to p in order to obtain:

$$\frac{\partial R'_{NS}(p)}{\partial p} = \frac{-B_{MM} \cdot \left(C_{mm}\left(m_{NS}^{ST}\right) - (n-p) \cdot B_{MM}\right) - (n-p) \cdot {B_{MM}}^2}{\left(C_{mm}\left(m_{NS}^{ST}\right) - (n-p) \cdot B_{MM}\right)^2} = \frac{-B_{MM} \cdot C_{mm}\left(m_{NS}^{ST}\right)}{\left(C_{mm}\left(m_{NS}^{ST}\right) - (n-p) \cdot B_{MM}\right)^2} > 0.$$

A.3 Proof of Proposition 1

We want to prove $M^{NC*}(p) > M^{ST*}(p)$ for every p, $1 \le p < n$. Let us assume the opposite, namely: $M^{NC*}(p) < M^{ST*}(p)$. We have $B_{MM} < 0$. Therefore, $R'_{NS} < 0$. Using the first order conditions in Table 1 and the general assumptions of the model, we have:

$$C_m\left(m_S^{ST*}\right) = p \cdot \left[B_M\left(M^{ST*}\right) \cdot \left(1 + R'_{NS}\right)\right] < p \cdot \left[B_M\left(M^{NC*}\right) \cdot \left(1 + R'_{NS}\right)\right] <$$

$$p \cdot \left[B_M\left(M^{NC*}\right)\right] = C_m\left(m_S^{NC*}\right)$$

for signatories and

$$C_m\left(m_{NS}^{ST*}\right) = B_M\left(M^{ST*}\right) < B_M\left(M^{NC*}\right) = C_m\left(m_{NS}^{NC*}\right)$$

for non-signatories. Hence, $C_m\left(m_S^{ST*}\right) < C_m\left(m_S^{NC*}\right)$, $C_m\left(m_{NS}^{ST*}\right) < C_m\left(m_{NS}^{NC*}\right)$. Therefore, given the convexity of cost functions, $m_S^{ST*} < m_{NS}^{NC*}$ and $m_{NS}^{ST*} < m_{NS}^{NC*}$ must hold. Hence, $M^{NC*}(p) > M^{ST*}(p)$, which contradicts our initial assumption $M^{NC*}(p) < M^{ST*}(p)$. Thus, we have: $M^{NC*}(p) > M^{ST*}(p)$. Consequently, $m_{NS}^{NC*}(p) < m_{NS}^{ST*}(p)$ must hold from the first order conditions of non-signatories and, consequently, $m_S^{NC*}(p) > m_S^{ST*}(p)$ must be true for $M^{NC*}(p) > M^{ST*}(p)$.

A.4 Proof of Proposition 2

a. Comparison with Nash equilibrium mitigation levels

Recall that for $p = \hat{p}$ we have $\hat{p} \cdot \left(1 + R'_{NS}(\hat{p})\right) = 1$. Moreover, the first order conditions of signatories (4) and those of non-signatories (5) in Table 1 are identical and equal to those in a Nash equilibrium, i.e., $B_M\left(M^{NE*}\right) = C_m\left(m^{NE*}\right)$. Hence, $m_S^{ST*}\left(\hat{p}\right) = m_{NS}^{ST*}\left(\hat{p}\right) = m^{NE*}$ and $M^{ST*}\left(\hat{p}\right) = M^{NE*}$.

Suppose $p < \hat{p}$. We want to show $M^{ST*}(p) < M^{NE*}$. By contradiction, we assume $M^{ST*}(p) > M^{NE*}$. From the first order conditions and knowing that $m_S^{ST*}(p) < m_{NS}^{ST*}(p)$ if $p < \hat{p}$, we have:

$$C_m\left(m_S^{ST*}\right) < C_m\left(m_{NS}^{ST*}\right) = B_M\left(M^{ST*}\right) < B_M\left(M^{NE*}\right) = C_m\left(m^{NE*}\right)$$

Therefore, given the convexity of cost functions, $m_S^{ST*} < m^{NE*}$ and $m_{NS}^{ST*} < m^{NE*}$ must hold. Hence, $M^{ST*}(p) < M^{NE*}$, which contradicts our initial assumption $M^{ST*}(p) > M^{NE*}$. Thus $M^{ST*}(p) < M^{NE*}$ must hold. Consequently, $m_{NS}^{ST*}(p) > m^{NE*}$ follows from the first order conditions of non-signatories and for $M^{ST*}(p) < M^{NE*}$ to hold, we must have $m_S^{ST*}(p) < m^{NE*}$. Hence, for $p < \hat{p}$, we have established: $m_S^{ST*}(p) < m^{NE*}$, $m_{NS}^{ST*}(p) > m^{NE*}$ and $M^{ST*}(p) < M^{NE*}$. For $p > \hat{p}$, the same kind of reasoning can be applied to establish: $m_S^{ST*}(p) > m^{NE*}$, $m_{NS}^{ST*}(p) < m^{NE*}$ and $M^{ST*}(p) > M^{NE*}$.

b. Comparison with Nash equilibrium welfare levels

For $p = \hat{p}$, it is obvious to conclude $W_S^{ST*}\left(\hat{p}\right) = W_{NS}^{ST*}\left(\hat{p}\right) = W_i^{NE*}$, $W^{ST}\left(\hat{p}\right) = W^{NE*}$. For non-signatories, we have $W_{NS}^{ST*}(p) < W_i^{NE*}$ if $p < \hat{p}$ and $W_{NS}^{ST}(p) > W_i^{NE*}$ if $p > \hat{p}$. This follows from the previous conclusions regarding total and individual mitigation levels. (Non-signatories have lower benefits and higher mitigation cost compared to the Nash equilibrium.)

For signatories, $W_S^{ST^*}(p) \geq W_i^{NE^*}$ follows from three pieces of information. 1) For $p=1$, $W_S^{ST^*}(1) > W_i^{NE^*}$ follows axiomatically. 2) Below, we show that $\frac{\partial W_S^{ST^*}(p)}{\partial p} <,=,> 0$ if $m_S^{ST^*} - m_{NS}^{ST^*} <,=,> 0$. 3) For $1 \leq p < \hat{p}$, $m_S^{ST^*} - m_{NS}^{ST^*} < 0$, for $p = \hat{p}$, $m_S^{ST^*} - m_{NS}^{ST^*} = 0$ (and $W_S^{ST^*}(\hat{p}) = W_i^{NE}$) and for $p > \hat{p}$, $m_S^{ST^*} - m_{NS}^{ST^*} > 0$.

For global welfare, we show first that $W^{ST^*}(p) < W^{NE^*}$ for $p < \hat{p}$. For $p < \hat{p}$, we know that $M^{ST^*}(p) < M^{NE^*}$ and $m_S^{ST^*}(p) < m_{NS}^{ST^*}(p)$. Let us a consider a hypothetical situation where we allocate total mitigation $M^{ST^*}(p)$ cost effectively such that $\ddot{m}_i = \frac{M^{ST^*}(p)}{n}$ due to the symmetry of strictly convex cost functions. We notice that $\ddot{m}_i < m_i^{NE^*}$, as we know $M^{ST^*}(p) < M^{NE^*}$. Moreover, we know that in a (symmetric) Nash equilibrium, M^{NE^*} is provided cost-effectively. Let $W_{CE}(m_i) = n \cdot \left[B_i(n \cdot m_i) - C_i(m_i) \right]$ be the cost-effective total welfare for a symmetric allocation of mitigation levels. Differentiating $W_{CE}(m_i)$ twice, gives: $n\left[n^2 \cdot B_{MM} - C_{mm} \right] < 0$, with the maximum at the socially optimal mitigation level $m_i^{SO^*}$, with $m_i^{NE^*} < m_i^{SO^*}$. Since $W_{CE}(m_i)$ is strictly concave, it follows that $W^{ST^*}(p) < \ddot{W} < W^{NE^*} < W^{SO^*}$.

For $p > \hat{p}$, $W^{ST^*}(p) > W^{NE^*}$ follows immediately, as $W_S^{ST^*}(p) \geq W_i^{NE^*}$ and $W_{NS}^{ST}(p) > W_i^{NE^*}$ as shown above.

c. Properties of mitigation levels

We investigate the sign of $\frac{\partial M^{ST^*}}{\partial p}$, $\frac{\partial m_S^{ST^*}}{\partial p}$ and $\frac{\partial m_{NS}^{*}}{\partial p}$ treating p as a continuous variable. Totally differentiating the first order conditions (4) and (5) in Table 1, rearranging terms and assuming third derivatives to be zero, we find:

$$\frac{\partial m_S^{ST^*}}{\partial p} = \frac{p \cdot B_{MM} \cdot \frac{\partial M^{ST^*}}{\partial p} \cdot \left(1 + R_{NS}^{'}\right)}{C_{mm}\left(m_S^{ST^*}\right)} + \frac{B_M \cdot \left(1 + R_{NS}^{'}\right)}{C_{mm}\left(m_S^{ST^*}\right)} \tag{A.1}$$

$$\frac{\partial m_{NS}^{ST^*}}{\partial p} = \frac{B_{MM} \cdot \frac{\partial M^{ST^*}}{\partial p}}{C_{mm}\left(m_{NS}^{ST^*}\right)}. \tag{A.2}$$

and, using, $\frac{\partial M^{ST*}}{\partial p} = m_S^{ST*} + p \cdot \frac{\partial m_S^{ST*}}{\partial p} - m_{NS}^{ST*} + (n-p) \cdot \frac{\partial m_{NS}^{ST*}}{\partial p}$ and substituting $\frac{\partial m_S^{ST*}}{\partial p}$ and $\frac{\partial m_{NS}^{ST*}}{\partial p}$ from above, we obtain:

$$\frac{\partial M^{ST*}}{\partial p} = \frac{m_S^{ST*} - m_{NS}^{ST*} + \frac{p \cdot B_M \cdot (1 + R'_{NS})}{C_{mm}(m_S^{ST*})}}{1 - B_{MM} \cdot \left[\frac{p^2 \cdot (1 + R'_{NS})}{C_{mm}(m_S^{ST*})} + \frac{(n-p)}{C_{mm}(m_{NS}^{ST*})} \right]} \tag{A.3}$$

We note that the term $\frac{B_M \cdot (1 + R'_{NS})}{C_{mm}(m_S^{ST})}$ in (A.1), $\frac{p \cdot B_M \cdot (1 + R'_{NS})}{C_{mm}(m_S^{ST})}$ as well as the denominator (which can be written as $1 - G'(M^{ST})$; see Appendix A.1) in (A.3) are positive. Further we know that $m_S^{ST} - m_{NS}^{ST} <,=,>$ for $p >,=,< \hat{p}$. Hence, we can immediately follow that in the range $\hat{p} \le p < n$, $\frac{\partial m_{NS}^{ST*}(p)}{\partial p} < 0$ and $\frac{\partial M^{ST*}(p)}{\partial p} > 0$. $m_S^{ST*}(p)$ must increase in p in some segment of this range, simply because $m_i^{NE*} = m_S^{ST*}(\hat{p}) < m_S^{ST*}(n) = m_i^{SO*}$. In the range $1 \le p < \hat{p}$, $m_{NS}^{ST*}(1) > m_i^{NE*}$, $m_S^{ST*}(1) < m_i^{NE*}$ and $M^{ST*}(1) < M^{NE*}$ is readily proved and $m_{NS}^{ST*}(\hat{p}) = m_i^{NE*}$, $m_S^{ST*}(\hat{p}) = m_i^{NE*}$ and $M^{ST*}(\hat{p}) = M^{NE*}$, as we know from above. Hence, we must have: $m_{NS}^{ST*}(p)$ decreases, $M^{ST*}(p)$ increases and $m_S^{ST*}(p)$ increases in p in some segment of this range.

d. Properties of welfare functions

Differentiating a non-signatory's welfare function with respect to p, we obtain:

$$\frac{\partial W_{NS}^{ST*}}{\partial p} = B_M \cdot \left[\frac{\partial M^{ST*}}{\partial p} \cdot \left(1 - \frac{B_{MM}}{C_{mm}(m_{NS}^{ST*})} \right) \right]$$

noting that $B_M > 0$ by assumption and $\left(1 - \frac{B_{MM}}{C_{mm}(m_{NS}^{*})} \right) > 0$ as $B_{MM} < 0$. Thus, the sign depends on the sign of $\frac{\partial M^{ST*}}{\partial p}$.

Differentiating signatories' welfare function with respect to p, we obtain: $\frac{\partial W_S^{ST*}}{\partial p} = B_M \cdot \left[\frac{\partial M^{ST*}}{\partial p} - p \cdot \left[1 + R'_{NS} \right] \cdot \frac{\partial m_S^{ST*}}{\partial p} \right]$ or

$$\frac{\partial W_S^{ST*}}{\partial p} = B_M \cdot \left[m_S^{ST*} - m_{NS}^{ST*} + (n-p) \cdot \frac{\partial m_{NS}^{ST*}}{\partial p} - p \cdot R'_{NS} \cdot \frac{\partial m_S^{ST*}}{\partial p} \right].$$

Plugging in $\frac{\partial m_{NS}^{ST*}}{\partial p}$, $\frac{\partial m_S^{ST*}}{\partial p}$ and $\frac{\partial M^{ST*}}{\partial p}$ from above, as well as $R'_{NS} = \frac{(n-p) \cdot B_{MM}}{C_{mm}\left(m_{NS}^{ST}\right) - (n-p) \cdot B_{MM}}$, we obtain:

$$\frac{\partial W_S^{ST*}}{\partial p} = B_M \left[\frac{\left(m_S^{ST*} - m_{NS}^{ST*} \right) \cdot C_{mm}\left(m_{NS}^{ST} \right)}{C_{mm}\left(m_{NS}^{ST} \right) - (n-p) \cdot B_{MM}} \right]$$

The denominator is positive as $B_{MM} < 0$. Thus, the sign of $\frac{\partial W_S^{ST*}}{\partial p}$ depends on the sign of $m_S^{ST} - m_{NS}^{ST}$ which in turn depends whether $p >,=,< \hat{p}$.

Finally, for total welfare, we use two pieces of information. First, we know $\frac{\partial W_{NS}^{ST*}}{\partial p} > 0$ if $\frac{\partial M^{ST*}}{\partial p} > 0$ for which a sufficient condition is $p \geq \hat{p}$. (This property is referred to as positive externality property.) Second, we establish a property called superadditivity for every p, $2 \leq p \leq n$, which is defined as follows: $p \cdot W_S^{ST*}(p) > [p-1] \cdot W_S^{ST*}(p-1) + W_{NS}^{ST*}(p-1)$. Hence, a sufficient condition for $\frac{\partial W^{ST*}}{\partial p} > 0$ is $p \geq \hat{p}$. (This property has been called full cohesiveness.) That is, the joint properties positive externality and superadditivity are sufficient properties conditions for full cohesiveness.

In order establish superadditivity, consider the following thought experiment in two steps. First consider any coalition enlargement from size $p-1$ to p. This implies one more signatory. Keeping total mitigation of the p signatories at the same level than at $p-1$ $\left(p \cdot m_S^{ST*}(p) = [p-1] m_S^{ST*}(p-1) + m_{NS}^{ST*}(p-1) \right)$, total mitigation cost will have decreased among the p signatories, as their aggregate mitigation is now shared with one more country. For the $n-p$ non-signatories nothing has changed. Second, the p Stackelberg leaders choose their

equilibrium strategies by maximizing their aggregate welfare, controlling the best-response of non-signatories. If they choose to change their strategies, their aggregate welfare must further increase. For the final move from $p-1=n-1$ to $p=n$, aggregate welfare also strictly increases as global welfare reaches its maximum at the social optimum $p=n$. Hence, under the ST-scenario, superadditivity holds for every p, $2\le p\le n$.

Hence, from above, it is immediately clear that in the range $\hat{p}\le p\le n$, $\frac{\partial W_{NS}^{ST*}(p)}{\partial p}>0$, $\frac{\partial W_{S}^{ST*}(p)}{\partial p}\ge 0$ and $\frac{\partial W^{ST*}(p)}{\partial p}>0$. In the range $1\le p<\hat{p}$, $W_{NS}^{ST*}(1)<W_{i}^{NE*}$, $W_{S}^{ST*}(1)>W_{i}^{NE*}$ and $W^{ST*}(1)<W^{NE*}$ is readily proved, and from above we have: $W_{NS}^{ST*}(\hat{p})=W_{i}^{NE*}$, $W_{S}^{ST*}(\hat{p})=W_{i}^{NE*}$ and $W^{ST*}(\hat{p})=W^{NE*}$. Hence, we must have: $W_{NS}^{ST*}(p)$ and $W^{ST*}(p)$ increase and $W_{S}^{ST*}(p)$ decreases in p in some segment of this range.

A.5 Proof of Proposition 3

We only provide a sketch of the proof and refer the interested reader to Online Appendix 2.

1) We know that any coalition of size $p\in[1,\hat{p}+1]$ is internally stable from Corollary 3 for our general welfare function (1). Hence, this also holds for welfare function (13).

2) We also know from Corollary 3 that if $\hat{p}+1\ge n$, then $p^{*}=n$.

3) Hence, we need to consider $\hat{p}+1<n$, which implies $\gamma>\frac{1}{n[n-2]}$ as established in (14) in the text.

4) For payoff function (13), it turns out that $\Omega(p)=-\frac{1}{2}\cdot\frac{a^2\cdot b^2\cdot c\cdot S}{T\cdot U^2}$ with

$T=(n-p)^2\cdot b^2+(p^2+2n-2p)\cdot n\cdot b\cdot c+n^2\cdot c^2>0$,

$U=((n-p)^2+2n-2p+1)\cdot b^2+(p^2+2n-4p+3)\cdot n\cdot b\cdot c+n^2\cdot c^2$ and

$S=\Phi_1\cdot b^4+n\cdot c\cdot\Phi_2\cdot b^3+n^2\cdot c^2\cdot\Phi_3\cdot b^2+n^3\cdot c^3\cdot\Phi_4\cdot b+n^4\cdot c^4\cdot(p^2-4p+3)$ where

$\Phi_1=-\left[(n-p)^4+(n-p)^3+(n-p)^2\right]$,

$$\Phi_2 = -p^4 + (2n+4)\cdot p^3 - (n^2+8n+2)\cdot p^2 + (6n^2+2n-2)\cdot p - 2n^3 - n^2 + 2n + 1,$$

$$\Phi_3 = p^4 - (2n+4)\cdot p^3 + (n^2+8n+9)\cdot p^2 - (4n^2+12n+12)\cdot p + 2n^2 + 8n + 5 \text{ and}$$

$$\Phi_4 = p^4 - 6p^3 + (2n+15)\cdot p^2 - (8n+18)\cdot p + 6n + 8.$$

If $\Omega(p) \geq 0$, then the coalition of size p is internally stable. The sign of $\Omega(p) \geq 0$ only depends on the sign of term S. Hence, we require $S \leq 0$ for internal stability.

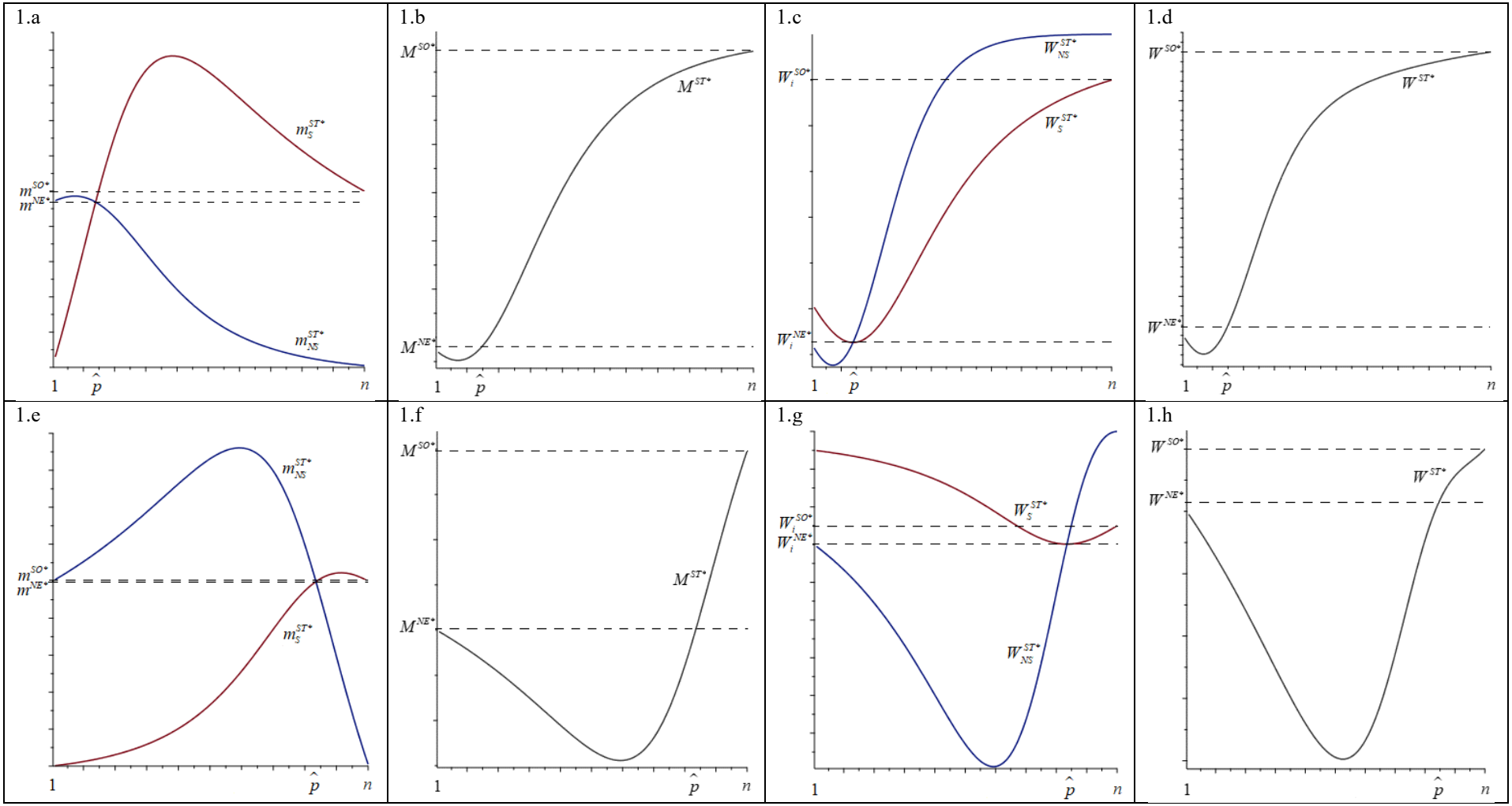
5) It can be shown that $S(\hat{p}+2)$ is a strictly convex function in $\gamma = c/b$. One can write $S(\hat{p}+2) = \gamma \cdot X(\gamma) - 4$ with $X(\gamma) = aa(n)\cdot\gamma^3 + bb(n)\cdot\gamma^2 + cc(n)\cdot\gamma + dd(n)$. It can be shown that $aa(n)$, $bb(n)$ and $cc(n)$ are strictly positive. Thus, it is obvious that $S(\hat{p}+2)$ is strictly increasing and strictly convex in γ. Hence, we insert the lowest possible γ, which is $\gamma = \frac{1}{n[n-2]}$, in $S(\hat{p}+2) = \gamma \cdot X(\gamma) - 4$. We find $S\left(\hat{p}+2, \gamma = \frac{1}{n[n-2]}\right) > 0$.

5) In principle, we cannot rule out that for $\hat{p}+2 < \hat{p}+x$ with $2 < x \leq n-1$ (we use the lower bound of $\hat{p}$, which is $\hat{p}=1$, as a conservative estimate and hence need to test for the highest possible value x at $\hat{p}+x \leq n$), we may have internal stability, i.e., $S(\hat{p}+x) \leq 0$. We restrict the analysis to $\hat{p}+1 < \hat{p}+x \leq n$ due to $\hat{p}+1 \geq n$ inducing $p^* = n$. $\hat{p}+x \leq n$ implies $\gamma \geq \frac{x}{n(n-x-1)}$.

6) It can be shown that $S(\hat{p}+x)$ is a strictly convex function in γ. We can rewrite $S(\hat{p}+x)$ as follows $S(\hat{p}+x) = \gamma \cdot Z(\gamma) - x^2(x-1)^2$ with $Z(\gamma) = ax(x,n)\cdot\gamma^4 + bx(x,n)\cdot\gamma^3 + cx(x,n)\cdot\gamma^2 + dx(x,n)\cdot\gamma + ex(x,n)$. It can be shown that $ax(x,n)$, $bx(x,n)$, $cx(x,n)$ and $dx(x,n)$ are positive. Thus, it is obvious that $Z(\gamma)$ is a strictly increasing and strictly convex function in γ. By inserting the lowest possible value for γ in $S(\hat{p}+x)$, which is $\gamma = \frac{x}{n(n-x-1)}$, we find $S\left(\hat{p}+x, \gamma = \frac{x}{n(n-x-1)}\right) = \frac{x(n-1)^5 \cdot f(x,n)}{(n-x-1)^5}$. Thus,

$sign\left[S\left(\hat{p} + x, \gamma = \frac{x}{n(n-x-1)} \right) \right] = sign\left[f(x,n) \right]$. For $2 \le x < n$, it can be shown that $f(x,n) > 0$. Hence, $S\left(\hat{p} + x, \gamma = \frac{x}{n(n-x-1)} \right) > 0$.

Figure 1: Mitigation and welfare levels as a function of coalition size p #



Plots are based on payoff function (13) introduced in section 4. We consider $n=100$ countries. For the first set of plots (plots 1.a to 1.d) $a=1$, $b=15$ and $c=1$ are assumed. For the second set of plots (plots 1.e to 1.h), $a=1$, $b=500$ and $c=1$ are assumed.