The Efficacy of International Environmental Agreements when Adaptation Matters: Nash-Cournot vs Stackelberg Leadership*

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 and Department of Environmental Sciences, Informatics and Statistics and Università Ca' Foscari Venezia, Italy

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

Abstract

We analyze the paradox of cooperation, as established by Barrett (1994), and later reiterated by many others, in a more general framework. That is, we show that stable coalitions are either small or if they are large, the potential gains from cooperation are small. First, we argue that the extension to a mitigation-adaptation game is a generalization of Barrett's pure mitigation game. Second, we consider for this extension not only the Nash-Cournot scenario, as in Bayramoglu et al. (2018), but also the Stackelberg scenario. Third, we show generally that if mitigation levels in different countries are strategic substitutes, stable coalitions are larger in the Stackelberg than in the Nash-Cournot scenario. Fourth, this is reversed if mitigation levels are strategic complements, which is possible if the strategic interaction between mitigation and adaptation is sufficiently strong. Fifth, for all possible combination of assumptions, we demonstrate that the paradox of cooperation is robust, except if mitigation and adaptation were strategic complements, which we argue is an assumption not supported by empirical evidence.

Keywords: Climate change, mitigation-adaptation game, international environmental agreements, paradox of cooperation, Nash-Cournot versus Stackelberg scenario

* We would like to thank Prof. Jean-Francois Jacques for very helpful comments on an earlier draft. Moreover, Michael Finus would like to acknowledge that many aspects of this paper have greatly benefitted from discussions with Prof. Santiago Rubio. This version has greatly benefitted from comments of three anonymous reviewers and the editor Prof. Andreas Lange on two previous versions of this paper.

1. Introduction

Mitigation and adaptation are two strategies to combat climate change. Mitigation directly targets at solving the cause of the problem, reducing greenhouse gas emissions, causing global warming. In contrast, adaptation aims at ameliorating the negative consequence of global warming. Whereas mitigation is typically viewed as a pure public good, adaptation is seen as a private good (reducing only damages of the party conducting adaptation). Addressing global warming requires international cooperation: isolated actions will not make a big difference if other countries do not follow suit. However, the signature and ratification of effective international climate agreements have proved to be difficult in the past. There is a widespread consensus that the Kyoto Protocol has not been able to curb the increase of greenhouse gases in the past, and also most scholars have doubts about the effectiveness of the Paris Accord signed in 2015, as highlighted by the latest IPCC 1.5 degrees report (IPCC 2018). As the effects of global warming become more and more visible, adaptation becomes increasingly important as a policy option. This is not only evident by the increasing literature on the costs and effectiveness of adaptation as well as about the practical and technical obstacles of implementation, in particular, in developing countries (IEG 2013 and World Bank 2010), but adaptation is also an integral part of almost all recent climate change negotiations (UNFCCC 2014 and 2016). The main obstacle of addressing the cause of global warming is the public good nature of mitigation. Reducing emissions comes at a cost that is borne by individual countries, but the benefits are enjoyed by all countries worldwide.

International climate negotiation failures have been largely explained by game-theoretic models of international environmental agreements (IEAs).[1] In the standard workhorse model with only

[1] The first models go back to Barrett (1994), Carraro and Siniscalco (1993) and Hoel (1992). This literature on IEAs has grown substantially over recent years. A collection of the most influential articles has been collected in a volume in Finus and Caparros (2015). In this volume, various extensions of the standard model are included for which in some cases more positive results are obtained. The importance of this topic is also highlighted by some recent papers, e.g., Battaglini and Harstad (2016) and Harstad (2012).

mitigation and symmetric players, only small agreements are stable if signatories and non-signatories choose their mitigation levels simultaneously, which has been called the Nash-Cournot scenario.[2] For Stackelberg leadership of signatories, more optimistic results have been obtained in terms of the size of stable agreements (Barrett 1994; Diamantoudi and Sartzetakis 2006; Rubio and Ulph 2006). However, as Barrett (1994) coined it, the paradox of cooperation persists: stable coalitions are either small or if they are large, the potential gains from cooperation are small. Recently, Bayramoglu et al. (2018) argued for the Nash-Cournot scenario that more optimistic results may be obtained if countries have a second strategy at their avail, namely adaptation, which they coined the mitigation-adaptation game. They show if the ***cross effect between mitigation and adaptation*** is sufficiently large, reaction functions in mitigation space may become upward sloping, associated with large stable agreements, including the grand coalition. This result does ***not*** depend on whether mitigation and adaptation are assumed to be substitutes (as commonly believed) or complements (as an unlikely possibility), but only that the rate of substitution or complementarity is large in absolute terms. However, Bayramoglu et al. (2018) neither measure the effectiveness of stable agreements nor do they consider Stackelberg leadership as an alternative assumption.

We acknowledge in this paper that mitigation-adaptation game is a generalization of Barrett's pure mitigation game for which we want to find out whether the paradox of cooperation is still a robust conclusion. We consider for this generalization not only the Nash-Cournot scenario, as in Bayramoglu et al. (2018), but also the Stackelberg scenario, as proposed by Eisenack and Kähler (2016) and Marrouch and Chaudhuri (2011). We show generally (neither resorting to specific payoff functions nor simulations) that if mitigation levels in different countries are strategic substitutes, stable coalitions are larger in the Stackelberg than in the Nash-Cournot scenario. This is reversed if mitigation levels are strategic complements, which is possible if the strategic interaction between mitigation and adaptation is sufficiently strong. For all possible combination of assumptions, we

[2] An exception is Karp and Simon (2013) who consider non-standard mitigation cost functions.

demonstrate that the paradox of cooperation is robust, except if mitigation and adaptation were strategic complements, which we argue is an assumption that lacks empirical support.

In what follows, we lay out the model in section 2, derive our general analytical results in section 3 and those for a specific payoff function in section 4. Section 5 evaluates the efficacy of stable agreements via extensive simulations and section 6 concludes with some hints about future research. All proofs are contained in the Appendix with further details provided in an Online Appendix.

2. The Model

2.1 Payoff Function

We consider n symmetric countries $i=1,2,...,n$, with N the set of all countries. Following Bayramoglu et al. (2018), the payoff function of every country i is given by:

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

The individual payoff comprises benefits B_i, which are a function of total mitigation, $M=\sum_{i=1}^{n}m_i$, and individual adaptation, a_i, minus the costs of mitigation C_i, which are a function of individual mitigation m_i and minus the costs of adaptation, D_i, which are a function of individual adaptation a_i. Benefits are a function of both strategies, total mitigation M and individual adaptation a_i. Both, mitigation, the pure public good, as well as adaptation, the pure private good, contribute to benefits.[3]

[3] It is generally known that the public good provision game can be alternatively framed as an emission game; they are dual problems. In the context of mitigation and adaptation, this is evident by comparing Bayramoglu et al. (2018) and Rubio (2018). In the emission game, the equivalent to the benefit function in the public good game is the damage function with aggregate emissions and adaptation being the arguments in this function. The importance of a correct conversion of mitigation into emission games, including possible problems, is discussed in Diamantoudi and Sartzetakis (2006) and Rubio and Ulph (2006), though without adaptation.

The strategy space of country i is given by $m_i \in \left[0, \overline{m_i}\right]$ and $a_i \in \left[0, \overline{a_i}\right]$. If we set $a_i = 0$ and assume $W_i\left(M, m_i, a_i = 0\right) = W_i\left(M, m_i\right)$, the pure mitigation game without adaptation can be retrieved. Thus, the mitigation-adaptation game can be viewed as a generalization of the mitigation game.

We assume that all countries have the same payoff function, i.e., all countries are assumed to be ex-ante symmetric. Hence, we can drop index i, whenever no misunderstanding is possible. However, as will become clear below, countries may nevertheless be ex-post asymmetric, as in our model countries endogenously choose whether they join an agreement and become signatories (S) or remain outside and become non-signatories (NS), and these groups choose different mitigation levels. If we want to stress this difference, we use subscript S and NS, respectively.

All welfare functions, as well as their first and second derivatives, are assumed to be continuous. Following Bayramoglu et al. (2018), we introduce the following assumptions where we denote for instance $B_M = {\partial B}/{\partial M}$, $B_{MM} = {\partial^2 B}/{\partial^2 M}$ and $B_{Ma} = B_{aM} = {\partial^2 B}/{\partial M \partial a_i}$.

General Assumptions

a) $B_M > 0,\ B_{MM} < 0,\ C_m > 0,\ C_{mm} > 0$.

b) $\lim\limits_{M \to 0} B_M > \lim\limits_{m \to 0} C_m > 0$.

c) $B_a > 0,\ B_{aa} \le 0,\ D_a > 0,\ D_{aa} \ge 0$.

If $B_{aa} = 0$, *then* $D_{aa} > 0$ *and vice versa: if* $D_{aa} = 0$, *then* $B_{aa} < 0$.

d) $\lim\limits_{a \to 0} B_a > \lim\limits_{a \to 0} D_a > 0.$

e) i) $B_{aM} = B_{Ma} < 0$ *or ii)* $B_{aM} = B_{Ma} > 0$.

These assumptions and their implications are discussed in Bayramoglu et al. (2018) where assumptions b) and d) are necessary conditions for an interior solution to which we confine ourselves in this paper. Mitigation and adaptation are substitutes, as commonly assumed for assumption e) i), and would be complements for assumption e) ii).[4] (See subsection 2.5 for a discussion.) It will become apparent that for almost all results the sign of the cross derivative does ***not*** matter, though the absolute value of this derivative will turn out to be important. In order to reduce the complexity of some of the subsequent proofs, we assume that third derivatives are equal to zero, which implies linear reaction functions. In the Appendix, we mention this assumption whenever we use it, but it will no longer be mentioned in the text.

2.2 The Coalition Formation Game

We consider the workhorse model of international environmental agreements, which is a two-stage cartel formation game. In the first stage, countries decide on their membership. Those countries, which join coalition P, $P \subseteq N$, are called signatories and those which remain outside are called non-signatories. In the second stage, signatories act as a single player, choosing their economic strategies by maximizing the aggregate payoff over all signatories. Non-signatories act as single players, maximizing their own payoff. The solution of the second stage leads to an economic strategy vector for every coalition P of size p, $1 \leq p \leq n$. If this strategy vector is unique, notation simplifies and we can write $W_i^*(p)$. As we will see below, as all signatories $i \in P$ choose the same strategy vector and the same applies to all non-signatories $j \notin P$ (though signatories and non-signatories will choose different strategy vectors) we can also write $W_S^*(p)$ and $W_{NS}^*(p)$, with the understanding that $W_{NS}^*(p)$ does not exist if $p = n$ and $W_S^*(p) = W_{NS}^*(p)$ if $p = 1$.[5] In Appendix 1, we provide a sufficient condition, which guarantees the existence and uniqueness of interior second stage equilibria.

For the second stage, we need to distinguish between the Nash-Cournot (NC) and the Stackelberg (ST) scenario. Under the NC-scenario, signatories and non-signatories choose their economic

[4] In the following, we rule out the uninteresting and special case of $B_{Ma} = 0$.

[5] Strictly speaking, $p = 0$ and $p = 1$ imply the same coalition structure. For notational simplicity, we assume $1 \leq p \leq n$. $W_S^*(p) = W_{NS}^*(p)$ if $p = 1$ for the Stackelberg scenario is explained below.

strategies simultaneously, and under the ST-scenario they do so sequentially, with signatories being the Stackelberg leader and non-signatories the followers, in line with the assumptions in the literature on IEAs (e.g., Barrett 1994 and Rubio and Ulph 2006).

Generally, if coalition P is empty ($p = 0$) or, which is equivalent, if it consists of only one player ($p = 1$), the equilibrium economic strategy vector will be the same as in the Nash equilibrium in games without coalition formation. This also means that we assume signatories can only assume Stackelberg leadership if $n > p > 1$ (but not if $p = 1$).[6] Conversely, if coalition $p = n$, i.e., the grand coalition has formed, this corresponds to the social optimum. There are no leaders and followers; hence, the NC- and ST-scenario coincide. Hence, difference in equilibrium strategies between the two scenarios in the second stage arise when there is partial cooperation, i.e., $1 < p < n$.

In the first stage, making already use of the symmetry assumption and the simplified notation because of a unique economic strategy vector for every coalition of size p, $1 \le p \le n$, a coalition of size p is stable if it is internally and externally stable.

$$\begin{aligned} &\text{Internal stability:} \qquad W_S^*(p) \ge W_{NS}^*(p-1) \\ &\text{External stability:} \qquad W_{NS}^*(p) \ge W_S^*(p+1)\ . \end{aligned} \tag{2}$$

Internal stability requires that a signatory has no incentive to leave a coalition of size p. External stability requires that a non-signatory has no incentive to join a coalition of size p. A coalition which is internally and externally stable is called stable and the size of such a coalition is denoted by p^*. It is important to note that despite second stage equilibria for $p = 1$ and $p = n$ are the same for the NC- and ST-scenario, internal stability for $p = n$ and external stability for $p = 1$ will be different.

[6] The alternative assumption of Stackelberg leadership also for $p = 1$ would only make a difference for stability at $p = 2$ which is anyway not very interesting for our subsequent analysis.

2.3 First Order Conditions and Slopes of the Reaction Functions in Mitigation Space and Mitigation-Adaptation Space

Under the NC-scenario, we assume in line with Bayramoglu et al. (2018) that all countries choose their mitigation and adaptation levels simultaneously. As shown by Bayramoglu et al. (2018), this is equivalent to all countries choosing first their mitigation levels and then all countries choosing their adaptation levels.

Under the ST-scenario, we assume that signatories simultaneously choose first their two economic strategies as leaders and then non-signatories do the same as followers. This is equivalent to any alternative sequence as long as signatories choose their mitigation levels first and each group does not choose adaptation before mitigation.[7]

In Table 1, we list the first order conditions in an interior equilibrium in the two alternative scenarios. (Second order conditions are provided in Appendix 1.) Consider first the NC-scenario. Signatories internalize the externality among its p members whereas non-signatories just maximize their own payoff. Hence, (3.a) and (3.b) imply $\frac{C_m\left(m_S^*\right)}{p} = C_m\left(m_{NS}^*\right)$ and therefore $m_S^* > m_{NS}^*$ due to the strict convexity of the mitigation cost function, where an asterisk indicates equilibrium values. According to (4), signatories and non-signatories will choose the same adaptation level in equilibrium, i.e., $a_i^* = a_S^* = a_{NS}^*$, as adaptation is a private good. Hence, $W_S^*(p) < W_{NS}^*(p)$ for any p, $1 < p < n$ (as all players have the same benefits and adaptation costs, but signatories have higher mitigation costs than non-signatories). Moreover, note that equilibrium adaptation only depends on total mitigation, i.e., $a_i^*(M)$, which is evident from (4).

Table 1 about here

[7] If adaptation was chosen before mitigation, the strategic role of adaptation would change and would lead to different outcomes (see Breton and Sbragia 2019, Eisenack & Kähler 2016 and Zehaie 2009) .

Let us now consider the ST-scenario. First, compared to the NC-scenario, it is evident from Table 1 that only the first order conditions of signatories regarding mitigation have changed. Second, again, $a_i^* = a_S^* = a_{NS}^*$ and $a_i^*(M)$. Third, we notice that the Stackelberg leaders choose their economic strategies such as to find the point on the followers' reaction function associated with the highest possible welfare for the leaders. That is, signatories as leaders take into consideration how non-signatories will react. Fourth, if we let $m_{NS} = r_{NS}(M_{-j})$ be the best-response of a non-signatory j, given the mitigation level of all other players except player j, M_{-j}, and using the symmetry assumption, which implies that all non-signatories de facto behave the same, we can define an aggregate best-response function of all non-signatories $M_{NS} = R_{NS}(M_S)$ with M_{NS} being the aggregate mitigation level of all non-signatories and M_S the aggregate mitigation level of all signatories. (Hence, $M = M_S + M_{NS}$.) Accordingly, $r_{NS}^{'}(M_{-j})$ and $R_{NS}^{'}(M_S)$ are the respective slopes of these best-response or reaction functions. Similarly, we can derive the slopes of individual and aggregate best-response functions of signatories, $r_S^{'}(M_{-i})$ and $R_S^{'}(M_{NS})$, with M_{-i} the total mitigation of all players except signatory i. See Bayramoglu et al. (2018), Proposition 2.

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

$$r_{NS}^{'}\left(M_{-j\notin P}\right) = \frac{\Psi}{C_{mm}\left(m_{NS}\right) - \Psi} \qquad R_{NS}^{'}\left(M_S\right) = \frac{\left(n-p\right) \cdot \Psi}{C_{mm}\left(m_{NS}\right) - \left(n-p\right) \cdot \Psi} \tag{8}$$

with $\Psi = B_{MM} + \dfrac{\left(B_{aM}\right)^2}{D_{aa} - B_{aa}}$.

In the mitigation-adaptation space, a country's reaction function is given by $a_i = h\left(M\right)$, with the associated slope given by

$$h'(M) = \frac{\partial a_i}{\partial M} = \frac{B_{aM}}{D_{aa} - B_{aa}}. \tag{9}$$

In mitigation space, reaction functions are downward sloping if $\Psi < 0$ and are upward sloping if $\Psi > 0$ (because the denominator of these slopes is positive if the second order conditions for a maximum hold, see Appendix 1). In mitigation-adaptation space, reaction functions are downward sloping if $B_{aM} < 0$ and upward sloping if $B_{aM} > 0$.[8]

Hence, reaction functions in mitigation space do not have to be downward sloping (as this would be the case in a game without adaptation) but can be upward sloping if adaptation is available as an additional strategy to mitigation. Thus, the leakage effect in terms of mitigation, due to mitigation levels in different countries being strategic substitutes, may turn into an anti-leakage effect such that mitigation levels become strategic complements. The latter possibility arises if $\Psi > 0$. An extensive discussion of this possibility is provided in the next two subsections. Important at this stage is to note that ***all*** our subsequent results only depend on the sign of Ψ, i.e., whether reaction functions in mitigation space are downward ($\Psi < 0$) or upward ($\Psi > 0$) sloping, but do ***not*** depend on whether mitigation and adaptation are substitutes or complements (i.e., the sign of $h'(M)$ does not matter).

With reference to Table 1, under the ST-scenario, comparing the first order conditions of signatories and non-signatories with respect to mitigation ((5.a) and (5.b)), we have $\frac{C_m\left(m_S^*\right)}{p \cdot \left(1 + R_{NS}'\right)} = C_m\left(m_{NS}^*\right)$. Hence, if $R_{NS}' > 0$, we conclude $m_S^*(p) > m_{NS}^*(p)$, given the convexity of the mitigation cost function and, because equilibrium adaptation levels of signatories and non-signatories are the same, $W_S^*(p) < W_{NS}^*(p)$ for any p, $1 < p < n$, follows. In contrast, if $R_{NS}' < 0$, $m_S^*(p) < m_{NS}^*(p)$ and, hence,

[8] We rule out the uninteresting and special case of $\Psi = 0$ and $B_{Ma} = 0$.

$W_S^*(p) > W_{NS}^*(p)$ is possible if p is small and/or if reaction functions are steep. For larger p and/or flat reaction functions, the reverse may hold: $m_S^*(p) > m_{NS}^*(p)$ and $W_S^*(p) < W_{NS}^*(p)$.

2.4 Technical Aspects of the Slopes of the Reaction Functions in Mitigation Space and Mitigation-Adaptation Space

If $\Psi = B_{MM} + \frac{(B_{aM})^2}{D_{aa} - B_{aa}} < 0$ reaction functions in mitigation space are downward sloping and if $\Psi = B_{MM} + \frac{(B_{aM})^2}{D_{aa} - B_{aa}} > 0$ they are upward sloping. We call $B_{MM} < 0$ the "direct effect" and $\frac{(B_{aM})^2}{D_{aa} - B_{aa}} > 0$ the "indirect effect", where the latter effect could also be called the "cross effect". The indirect effect is always positive because $D_{aa} - B_{aa} > 0$ from our general assumptions and the sign of B_{aM} does not matter because it is squared.[9] In a game without adaptation, the indirect effect does not exist and because $B_{MM} < 0$ from our General Assumptions, $\Psi < 0$ always holds. In a game with adaptation, $\Psi > 0$ is possible if $|B_{MM}| < \frac{(B_{aM})^2}{D_{aa} - B_{aa}}$. That is, "direct effect" must be smaller than the "indirect effect". In the terminology of Ebert and Welsch (2012, p.54), the indirect effect is the "adaptation capacity" of a country, with a high capacity favouring large values of Ψ. Apart from a strong interaction of mitigation and adaptation on the benefit side (i.e., large absolute values of

[9] The indirect effect can be broken down such that $\frac{(B_{aM})^2}{D_{aa} - B_{aa}} = B_{aM} \cdot \frac{\partial a_i}{\partial M} > 0$. The indirect effect is always positive because B_{aM} and $h^{'}(M) = \frac{\partial a_i}{\partial M}$ have always the same sign), namely either $B_{aM} < 0$ and $\frac{\partial a_i}{\partial M} < 0$ if mitigation and adaptation are substitutes, as commonly believed, or $B_{aM} > 0$ and $\frac{\partial a_i}{\partial M} > 0$ if mitigation and adaptation are complements, which is normally seen as an unrealistic assumption (see subsection 2.5). Therefore, for the sign of the slopes of the reaction functions in mitigation space, the sign of B_{aM} and $\frac{\partial a_i}{\partial M}$ does ***not*** matter. Only the absolute value of the indirect effect matters and for $\Psi > 0$, this indirect must be sufficiently strong compared to the direct effect. See Appendix 1 for further explanations.

B_{aM}), a large adaptation capacity is favoured by small values of $D_{aa} - B_{aa}$. Small values of D_{aa} imply a flat marginal adaptation cost function and small values of B_{aa} imply a "small extent by which the effectiveness of adaptation diminishes".

In order to understand the technical role of $\varPsi$ better, we adopt the concept of "minimized total costs" of Rubio (2018) for our purposes, which we call "optimally adapted net benefits", which are given by $B_i(M, a_i(M)) - D_i(a_i(M))$. That is, we consider that equilibrium adaptation is a function of total mitigation in the benefit function, correcting for the cost of adaptation. Differentiating the optimally adapted net benefit function with respect to mitigation M twice, gives $\varPsi$. Hence, the optimally adapted net benefit function is concave if $\varPsi < 0$ and convex if $\varPsi > 0$. However, as we show in Appendix 1, even if $\varPsi > 0$, the optimization problem of the entire payoff function is still concave if the second order condition for a maximum hold. For instance, for $p = 1$, the second order condition is given by $\varPsi - C_{mm}(m_i) < 0$. Thus, if $\varPsi > 0$, we need $\varPsi < C_{mm}(m_i)$ for this condition to be satisfied.[10, 11]

In order to illustrate the relation between mitigation and adaptation, let all mitigation cost functions be multiplied by a cost parameter c such that costs are given by $c \cdot C(m_i)$. Then from the first order conditions for mitigation, it is clear that ${\partial m_i^*}/{\partial c} < 0$ for all players and, hence, ${\partial M^*}/{\partial c} < 0$. This drop in total mitigation will transmit into a change of equilibrium adaptation through $h'(M) = \frac{\partial a_i}{\partial M}$, which implies that equilibrium adaptation will increase if mitigation and adaptation are substitutes and will decrease if they are complements. Thus, for upward sloping reaction functions in mitigation space,

[10] Non-convexity of damages in the context of multiple externalities has already been treated for instance in Baumol and Bradford (1972) and Starrett (1972).

[11] Further technical aspects are discussed in Appendix 1.

we either need a large rate of substitution or a large rate of complementarity between mitigation and adaptation.

Finally, let us ask the questions whether there could be any reason why mitigation and adaptation are complements, despite we adhere to the standard assumption of $B_{aM} < 0$. Clearly, in our model, $B_{aM} < 0$ implies $h'(M) = \frac{\partial a_i}{\partial M} < 0$. In Ingham et al. (2013) this is confirmed for most model versions which the authors consider. Nevertheless, they point to one exception, namely if adaptation costs depend on the level of mitigation in a particular form. Let $D(a, M)$. Then it is straightforward to show that $h'\left(M\right) = \frac{\partial a_i}{\partial M} = \frac{B_{aM} - D_{aM}}{D_{aa} - B_{aa}}$. Now if $B_{aM} < 0$, $h'\left(M\right) > 0$ requires $D_{aM} < 0$ and $\left|B_{aM}\right| < \left|D_{aM}\right|$. That is, mitigation reduces the marginal cost of adaptation and the cross effect on the cost function is larger than on the benefit function. The argument for $D_{aM} < 0$ could be that by reducing emissions, the production capacity of adaptation is enhanced. Even if one buys into this argument, it is likely that this does not hold for all levels of emissions, but only for very high levels of emissions above some threshold. That is, in our context, $D_{aM} < 0$ may hold for low levels of M but not for higher levels of M. Moreover, one may argue that if input markets are not perfectly competitive, then $D_{aM} > 0$ (as well as $C_{ma} > 0$) is a more reasonable assumption.

2.5 Empirical Aspects of the Slopes of the Reaction Functions in Mitigation and Mitigation-Adaptation Space

Empirical evidence about the slope of reaction functions in mitigation space is difficult to obtain. One could be inclined to look for empirical estimates from climate models, Computable General Equilibrium models (CGE) or Integrated Assessment Models (IAM). However, these models typically do not capture the strategic interaction between regions in a game-theoretic sense and even if they do, to the best of our knowledge, we have not found estimates about the slopes of the reaction functions. Moreover, the estimation of leakage effects as a proxy is also not very useful as simulations typically assume only a unilateral policy intervention.

There are quite some papers which econometrically test for the strategic interaction between countries or regions for different environmental problems (e.g., Fredriksson and Millimet 2002a, b, Grubb et al. 2002, Murdoch et al. 1997 and Perkin and Neumayer 2008, 2009; see Brueckner 2003 for an overview) of which some find a positive correlation between environmental standards in different countries/regions. Of course, not all of those studies are related to climate change and, even more important, adaptation does not play a role. Positive correlations in those papers are mainly driven by political and technological spillovers as well as trade. Political spillovers put pressure on neighboring political institutions to follow suit; technological spillovers reduce abatement costs in other regions and, hence, encourage the implementation of higher environmental standards; imports of advanced technology improve environmental standards as a by-product, in particular, in developing countries.

From an extensive screening of the literature about the relation between mitigation and adaptation, we have found only one instance, namely Yohe and Strzepek (2007), who claim to have found a complementary relationship for flood prevention measures along the Brahmaputra and Ganges rivers in India. The argument seems to be along the lines which we have presented in subsection 2.4: adaptation increases the productivity of mitigation and vice versa. However, we need to point out that the bulk of the empirical literature on adaptation only focuses on the benefits and costs of adaptation, point to the fact that mitigation will not be sufficient to address the climate change problem, but do not investigate the rate of substitution or complementarity between mitigation and adaptation.

3. General Results

In the following analysis, we focus on comparing the size and success of stable agreements under the NC- and ST-scenario. In this section, we derive some results based on the general payoff function (1). The derivation of equilibrium coalition sizes is dealt with in section 4, as this requires the assumption of a specific payoff function. Only finally, when it comes to the evaluation of the efficacy of stable coalitions, do we need to resort to simulations on which we report in section 5.

3.1 Definitions

In order to explain differences between the NC- and ST-scenario, it will be helpful to consider some general properties of the coalition formation game.

Definition 1: Positive Externality, Positive Internalisation, Superadditivity and Cohesiveness

Let $n \geq p \geq 2$.

i) *PEP: The expansion of coalition* $p-1$ *to* p *exhibits a positive (negative) externality to non-signatories if:*

$$W_{NS}^{*}(p) > (<) W_{NS}^{*}(p-1).$$

If this holds for all p, $n \geq p \geq 2$, *the game is a positive (negative) externality game.*

ii) *PIP: The expansion of coalition* $p-1$ *to* p *exhibits a positive (negative) internalisation to signatories if:*

$$W_{S}^{*}(p) > (<) W_{S}^{*}(p-1).$$

If this holds for all p, $n \geq p \geq 2$, *the game is a positive (negative) internalisation game.*

iii) *SAD: The expansion of coalition* $p-1$ *to* p *is superadditive if:*

$$p \cdot W_{S}^{*}(p) > [p-1] \cdot W_{S}^{*}(p-1) + W_{NS}^{*}(p-1).$$

If this holds for all p, $n \geq p \geq 2$, *the game is superadditive.*

iv) *WCOH: The expansion of coalition* $p-1$ *to* p *is welfare cohesive if:*

$$p \cdot W_{S}^{*}(p) + [n-p] \cdot W_{NS}^{*}(p) > [p-1] \cdot W_{S}^{*}(p-1) + [n-p+1] \cdot W_{NS}^{*}(p-1)$$

If this holds for all p, $n \geq p \geq 2$, *the game is welfare cohesive.*

v) *MCOH: The expansion of coalition* $p-1$ *to* p *is mitigation cohesive if:*

$$p \cdot M_{S}^{*}(p) + [n-p] \cdot M_{NS}^{*}(p) > [p-1] \cdot M_{S}^{*}(p-1) + [n-p+1] \cdot M_{NS}^{*}(p-1)$$

If this holds for all p, $n \geq p \geq 2$, the game is mitigation cohesive.

The first three properties may be viewed as positive properties in that they help to explain whether stable coalitions will be small or large. A positive externality makes it attractive to stay outside a coalition whereas for a negative externality just the opposite holds. Positive internalisation and superadditivity may be viewed as necessary conditions to make joining a coalition attractive. In a superadditive and negative externality game, the grand coalition is the unique stable agreement (Weikard 2009). Thus, cooperation does not pose a problem. In contrast, in positive externality games, stable coalitions are typically small. This is evident if the properties positive internalisation and superadditivity fail, but even if they hold, the positive externality effect to outsiders may be stronger than the positive externality effect to insiders via positive internalisation and superadditivity, such that only small coalitions are stable.[12]

The fourth and the fifth property can be viewed as normative properties. Clearly, in the grand coalition, total welfare and total mitigation levels are strictly higher than in any other coalition, which is true in any externality game. However, it may not always be true that these levels increase with every enlargement of a coalition, irrespective of its size, as we will illustrate and explain in more detail below. Note that a sufficient condition for welfare cohesiveness is superadditivity and positive externalities.

3.2 Propositions

Our first result is summarized in Proposition 1 below.

[12] Note that whenever we have $m_S(p) > m_{NS}(p)$ (which is always the case in the NC-scenario and in the ST-scenario if $\Psi > 0$; see subsection 2.3), positive internalisation and superadditivity when moving from $p-1$ to p are necessary, though not sufficient, properties for internal stability of a coalition of size p. See the proof in Bayramoglu et al. (2018), Appendix A.2.

Proposition 1: Comparison of the NC- and ST-Scenario, Mitigation, Payoffs and Stable Coalitions

Consider a generic coalition of size p .

a) *Suppose* $\Psi < 0$ *. Hence, reaction functions are downward sloping in mitigation space. Then the following relations hold:*

- $M^{NC*}(p) > M^{ST*}(p)$*,* $m_S^{NC*}(p) > m_S^{ST*}(p)$ *and* $m_{NS}^{NC*}(p) < m_{NS}^{ST*}(p)$ $\forall\ p, 1 < p < n$*;*
- $W_S^{NC*}(p) < W_S^{ST*}(p)$ *and* $W_{NS}^{NC*}(p) > W_{NS}^{ST*}(p)$ $\forall\ p, 1 < p < n$;
- $p^{ST*} \geq p^{NC*}$ $\forall\ p, 1 \leq p \leq n$*, with* $p^{ST*} \geq 2$.

b) *Suppose* $\Psi > 0$ *. Hence, reaction functions are upward sloping in mitigation space. Then the following relations hold:*

- $M^{NC*}(p) < M^{ST*}(p)$*,* $m_S^{NC*}(p) < m_S^{ST*}(p)$ *and* $m_{NS}^{NC*}(p) < m_{NS}^{ST*}(p)$ $\forall\ p,\ 1 < p < n$*;*
- $W_S^{NC*}(p) < W_S^{ST*}(p)$*,* $W_{NS}^{NC*}(p) < W_{NS}^{ST*}(p)$ *and* $W^{NC*}(p) < W^{ST*}(p)$ $\forall\ p,\ 1 < p < n$;
- $m_S^{ST*}(p) - m_S^{NC*}(p) > m_{NS}^{ST*}(p) - m_{NS}^{NC*}(p)$*, implying*
  $W_S^{ST*}(p) - W_S^{NC*}(p) < W_{NS}^{ST*}(p) - W_{NS}^{NC*}(p)$ $\forall\ p,\ 1 < p < n$*;*
- $p^{NC*} \geq 2$ and $p^{ST*} \geq 2$.

Proof: See Appendix A.2.

If $\Psi < 0$ (Proposition 1a), which would always be true in a pure mitigation game without adaptation, reaction functions in mitigation space are downward sloping. Consequently, signatories, having a strategic advantage (i.e., a first mover advantage) under the ST-scenario, will lower their mitigation levels compared to the NC-scenario, knowing that non-signatories will partly make up for this by mitigating more. Overall, total mitigation will be lower under the ST- than under the NC-scenario for any generic coalition of size p , $n > p > 1$. The Stackelberg leader will be better off and the reverse is true for the follower compared to the NC-scenario. It is for this reason that stable coalitions under the ST-scenario will be at least as large as under the NC-scenario. This result is known in the literature since Barrett (1994), though it has only been derived from simulations in the pure mitigation game. We provide a general proof, including the generalization to a mitigation-adaptation game.

It is evident from Proposition 1a why it is not possible to draw any general conclusion about total mitigation levels and global welfare for stable coalitions under the two scenarios. In terms of global welfare, we do not know whether $W^{NC*}\left(p^{NC}\right) > W^{ST*}\left(p^{ST}\right)$ or whether the reverse is true for a given p, as signatories are better off but non-signatories worse off under the ST- than under the NC-scenario. Hence, we also do not know generally whether for stable coalitions $W^{NC*}\left(p^{NC*}\right) < W^{ST*}\left(p^{ST*}\right)$ or whether the opposite is true in equilibrium. In terms of global mitigation, we know that $M^{NC*}\left(p\right) > M^{ST*}\left(p\right)$ but $p^{NC*} \leq p^{ST*}$ and, hence, generally, $M^{NC*}\left(p^{NC*}\right) <,> M^{ST*}\left(p^{ST*}\right)$.

If $\Psi > 0$ (Proposition 1b), reaction functions in mitigation space are upward sloping, which is only possible in a game which includes adaptation as a strategy to address climate change. Both, signatories and non-signatories, increase their mitigation levels under the ST- compared to the NC-scenario in such a matching game. This also translates into a Pareto-improvement to all countries and therefore in higher total welfare. However, compared to the NC-scenario, non-signatories gain more than signatories, i.e., there is a second mover advantage.[13] The reason is that signatories increase their mitigation levels more than non-signatories and therefore carry higher additional mitigation costs. Hence, one could be inclined to conclude that the size of stable coalitions is generally weakly smaller under the ST- than NC-scenario.[14] However, we have not been able to prove this at a general level, even though this is confirmed for the specific payoff function which we consider in section 4. Finally, even if we had always $p^{NC*} > p^{ST*}$, and know that $M^{NC*}\left(p\right) < M^{ST*}\left(p\right)$ as well as

[13] This is in line with the literature on Stackelberg games with symmetric players (though usually confined to two players). There is a first (second) mover advantage in the presence of downward (upward) sloping reaction functions (Endres 1992 and Gal-Or 1985).

[14] We would like to thank an anonymous reviewer for pointing out this potential pitfall. The analysis in Appendix A.5 underlines why general conclusions are not straightforward.

$W^{NC*}(p) < W^{ST*}(p)$ from Proposition 1b, nothing could be concluded about total mitigation and welfare for stable coalitions. This would only be possible for $p^{NC*} < p^{ST*}$ for which, however, we do not find evidence.

In order to understand the driving forces in the coalition game and to rationalize equilibrium coalition sizes as well as differences between the two scenarios, we consider some general properties in Proposition 2 below (see Definition 1 above). These will be particularly helpful in explaining our more specific results in section 4.

Proposition 2: Properties under the NC- and ST-scenario

Consider the general payoff function (1) and assume the General Assumptions to hold. Further assume the sufficient conditions for the existence of a unique interior equilibrium in the second stage to hold, as stated in Appendix 1. Then the following conclusion can be drawn:

Properties	$\Psi < 0$		$\Psi > 0$	
	NC-scenario	**ST-scenario**	**NC-scenario**	**ST-scenario**
PEP	✓	fails when MCOH fails	✓	✓
PIP	may fail for small p	may fail for small p; holds if $m_S^*(p) > m_{NS}^*(p)$	✓	✓
SAD	may fail for small p	✓	✓	✓
WCOH	may fail for small p	may fail for small p	✓	✓
$m_{NS}^*(p) - m_{NS}^*(p-1)$	-	- if MCOH holds; + possible if MCOH fails	+	+
$m_S^*(p) - m_S^*(p-1)$	- possible for large p	- possible for large p	+	+
MCOH	✓	may fail for small p	✓	✓

Properties as defined in Definition 1; ✓ = property holds for all expansion $p-1$ to p, $2 \le p \le n$, except for PEP for which $2 \le p \le n-1$.

Proof: See Appendix A.3.

Under the NC-scenario, the game is a positive externality game. Total mitigation increases steadily with an expansion of the coalition from which also non-signatories benefit due to the nonexclusiveness of the public good.[15] Non-signatories reduce their contribution to this public good if reaction functions are downward sloping (and therefore have not only higher benefits but also lower mitigation costs). However, even if $\Psi > 0$, non-signatories contribute less than proportionally to the total increase in total mitigation (see Proposition 1) and, hence, also enjoy a positive externality from the expansion of the coalition. Therefore, with positive externalities, there is an incentive to remain a non-signatory.

Moreover, under the NC-scenario, if $\Psi < 0$, it is also evident that positive internalisation and superadditivity may fail due to the leakage effect, which is also an obstacle to form large stable coalitions. In particular if p is small, there are many non-signatories countervailing the efforts of the few signatories. Together, this explains why only small coalitions are stable if reaction functions are downward sloping in mitigation space. This will be confirmed for our specific payoff function considered in section 4. In contrast, if reaction functions are upward sloping in mitigation space, positive internalisation and superadditivity always hold, as the game has turned into a matching game with anti-leakage. This allows to form larger stable coalitions, including the grand coalition if $\Psi > 0$, as confirmed for our specific payoff function in section 4. It is also evident that if the leakage effect is present (i.e., $\Psi < 0$), welfare cohesiveness may fail (as a result of a failure of superadditivity).

Under the ST-scenario, the negative conclusion about the size of stable coalitions if reaction functions are downward sloping (i.e., $\Psi < 0$) is just reversed. Roughly speaking, and as our simulations will confirm in section 5, the steeper the reaction function, the larger is the strategic advantage of the

[15] The reader may be puzzled how $\partial M^* / \partial p > 0$ can hold if $\partial m^*_{NS} / \partial p < 0$ holds and $\partial m^*_S / \partial p < 0$ is possible, where we treat p as a continuous variable without loss of generality. However, in Appendix A.3, we show $\partial M^*_S(p) \big/ \partial p > 0$, $\partial M^*_{NS}(p) \big/ \partial p < 0$ but $\partial M^*_S(p) \big/ \partial p > \left| \partial M^*_{NS}(p) \big/ \partial p \right|$ if $\Psi < 0$. Intuitively, even if an "old" signatory at $p-1$ may not increase its mitigation level at p, the "old" non-signatory at $p-1$ which is a "new" signatory at p increases its mitigation level and, hence, signatories as an entire group increase their mitigation level and this increase is larger than the decrease of total mitigation of non-signatories.

leader over the follower and, hence, the larger will be stable coalitions. Moreover, superadditivity always holds, and, at least for not too large coalitions, the enlargement of coalitions may not be associated with positive but with negative externalities, making it attractive for non-signatories to join the coalition. The fact that larger coalitions may not necessarily lead to substantially better outcomes, as will be confirmed in section 5, is already apparent by the fact that welfare and mitigation cohesiveness does not generally hold if $\Psi < 0$.[16] In other words, larger stable coalitions under Stackelberg leadership come at a price.

For $\Psi > 0$, under the ST-scenario all properties hold. In fact, at this level of generality, all properties are the same than under the NC-scenario, from which we may conjecture that stable coalitions are of similar size. Section 4 will confirm this for our specific payoff function.

4. Stable Coalitions for a Specific Payoff Function

4.1 Preliminaries

It is well-known from the literature on IEAs that sharp predictions about first stage equilibria (i.e., the size of stable agreements) are only possible for specific payoff functions. Different from the main body of the literature, we are able to provide analytical results. In line with the literature on IEAs and following Bayramoglu et al. (2018), we consider a payoff function with quadratic benefit and cost functions:

$$w_i = \left(bM - \frac{g}{2} M^2 \right) + a_i \left(\beta - fM \right) - \frac{c}{2} m_i^2 - \frac{d}{2} a_i^2 \text{ such that } B_{aM} < 0 \tag{10.a}$$

and

$$w_i = \left(bM - \frac{g}{2} M^2 \right) + a_i \left(\beta + fM \right) - \frac{c}{2} m_i^2 - \frac{d}{2} a_i^2 \text{ such that } B_{aM} > 0 \tag{10.b}$$

[16] Welfare cohesiveness fails whenever the superadditivity effect is dominated by the negative externality effect. Mitigation cohesiveness may fail as the Stackelberg leaders use their strategic advantage to reduce their contribution to the public good, which may not be compensated by the followers' additional mitigation effort.

assuming that all parameters b, g, c β, f, and d are strictly positive. If we were to set $g=0$, we could retrieve the linear-quadratic payoff function, also frequently considered in the literature on IEAs. For expositional clarity, we ignore this case. It is also clear that by setting $a_i=0$, we could retrieve the pure mitigation game. Due to space limitations, we provide all details about equilibrium mitigation and adaptation levels under the NC- and ST-scenario and the conditions that we impose on the parameters, such that the sufficient conditions for existence and uniqueness of interior equilibria are satisfied, in Appendix A.4.

Note that (10a) and (10b) only differ in one sign, which does ***not*** make any difference for the size of stable coalitions. In other words, for all results derived in this section, one can work with (10.a) and if the case of $B_{aM}>0$ shall be considered, then parameter f is set to $-f$. As the discussion here and in Appendix A.4 will confirm, the sign of f does not matter, as it always appears in the form f^x, with $x=2,4,6,....$, i.e., even numbers, in all relevant equations. Hence, all results are the same, regardless whether we assume $B_{aM}<0$ or $B_{aM}>0$.

For payoff function (10.a) and (10.b), $\Psi=\dfrac{f^2-g\cdot d}{d}$, which is negative if $f^2-g\cdot d<0$ and positive if $f^2-g\cdot d>0$, but the sign of parameter f does not matter. Accordingly, the slopes of the aggregate reaction function of non-signatories, $R_{NS}^{'}=\dfrac{(n-p)\left(f^2-g\cdot d\right)}{c\cdot d-(n-p)\left(f^2-g\cdot d\right)}$ and signatories, $R_S^{'}=\dfrac{p^2\left(f^2-g\cdot d\right)}{c\cdot d-p^2\left(f^2-g\cdot d\right)}$, are negative if $\Psi<0$ and positive if $\Psi>0$, as the denominators of these slopes are always positive due the sufficient condition for a unique interior equilibrium, which for payoff functions (10a) and (10b) is also a necessary condition and is given by $c\cdot d-n^2\left(f^2-g\cdot d\right)>0.$ Further details are provided in Appendix A.4.

Proposition 3: Stable Coalitions for Specific Welfare Functions

Consider payoff function (10.a) and (10.b) and assume the conditions imposed on the parameters in Appendix A.4 to hold.

The size of stable coalitions p^* under the NC- and ST-scenario are as follows:

Let $\Psi < 0$.

Under the NC-scenario, $p^{NC*} \in [1,2]$ and under the ST-scenario $p^{ST*} \in [2,n]$. Thus, $p^{NC*} \leq p^{ST*}$.

Let $\Psi > 0$ and $n \geq 7$.

a) *Under the NC- and ST-scenario, $p^* = 3$ is always a stable equilibrium.*

b) *If $\Psi \geq \underline{\Psi}$, then $p^* = n$ is a second equilibrium where $\underline{\Psi}^{NC} < \underline{\Psi}^{ST}$. That is, upward sloping reaction functions in mitigation space must be sufficiently steep such that the grand coalition is stable.*

c) *There are no other equilibria than $p^* = 3$ and $p^* = n$. If $p^* = n$ exists, it Pareto-dominates $p^* = 3$.*

d) *Thus, $p^{NC*} \geq p^{ST*}$.*

Proof: See Appendix A.5.

It is evident that for downward sloping reaction functions in mitigation space, under the ST-scenario, even the grand coalition could form. In contrast, under the NC-scenario, only small coalitions are stable. As we will see in the next subsection, the steeper reaction functions in mitigation space are, the larger will be stable coalitions in the ST-scenario. The steeper the reaction function, the larger is the strategic advantage of signatories, and the larger the strategic loss of non-signatories compared to the NC-scenario.

For upward sloping reaction functions in mitigation space, things are more complicated. First note that the upper bound of Ψ ($\overline{\Psi} = c/n^2$) follows for payoff function (10a) and (10b) from the conditions for the existence of a unique interior equilibrium, which are identical to the second order conditions in the grand coalition (see Appendix A.1 and A.4). Since the slopes of the reaction functions in mitigation space increase in Ψ, this imposes an upper bound on the maximum value of those slopes. Second, in the entire permissible range of Ψ a coalition of size $p=3$ is stable. Third, if Ψ is larger than some threshold ($\Psi \geq \underline{\Psi}$), which implies that the slopes of the reaction functions in mitigation are sufficiently large, also the grand coalition is stable. Fourth, this lower bound is larger under the ST- than under the NC-scenario. Consequently, there is a range of Ψ, $\Psi \in [\underline{\Psi}^{NC}, \underline{\Psi}^{ST})$, such that $p^{NC*} = n > p^{ST*}$, which specifically means $p^{NC*} = n > p^{ST*} = 3$ if $n \geq 7$. Fifth, if the grand coalition is stable, it Pareto-dominates $p^{*} = 3$.[17]

The intuition why the grand coalition can be stable if reaction functions in mitigation space are upward sloping ($\Psi > 0$) and sufficiently steep (large values of Ψ), can be related to two characteristics. First, according to Proposition 2, we know that for $\Psi > 0$ the properties superadditivity and positive internalization hold for any expansion of coalitions under both scenarios. As we explained above, both properties provide an incentive for countries to join a coalition. Second, normally, stable coalitions are small because if they are sufficiently large, leaving the coalition is attractive: leaving decreases (concave) benefits only marginally but implies a substantial drop in (convex) mitigation costs. We know from section 2.4 that for $\Psi > 0$ the optimally adapted net benefit function is convex. Particular for larger values of Ψ and/or large membership, this function is very steep. Hence, leaving a coalition at $p = n$ implies a large drop of convex optimally adapted net benefits, which may exceed the large drop of mitigation costs. Thus, leaving does not pay.

5. The Paradox of Cooperation

In this section, we want to analyse the "paradox of cooperation". For this, we need to evaluate stable coalitions in welfare terms. We consider two indices in our simulations. We recall that no-cooperation

[17] Bayramoglu et al. (2018) show that for welfare function (10.a) and (10.b) in the NC-scenario and $\Psi > 0$ $p^{NC*} \in [3, n]$. We find that if $n \geq 7$ (as also assumed in our simulations), this leads to $p^{NC*} = \{3, n\}$. See Appendix A.5.

with $p=1$ corresponds to the classical Nash equilibrium without coalition formation and full cooperation with $p=n$ corresponds to the social optimum. We denote total welfare by W, $W=\sum_{i=1}^{n} w_i$, and use superscripts to refer to the social optimum, SO, Nash equilibrium, NE, and stable coalitions in the NC- and ST-scenario, respectively.

Definition 2: Importance of Cooperation and Improvement upon No Cooperation

- *The Importance of Cooperation Index (ICI) measures the percentage global welfare improvement from moving from no-cooperation (NE) to the social optimum (SO):*

$$ICI = \frac{W^{SO} - W^{NE}}{W^{NE}} \cdot 100$$

- *The Improvement upon the No Cooperation Index (INI) measures the percentage global welfare improvement obtained in a stable equilibrium under the NC- and ST-scenario, respectively:*

$$INI^{NC} = \frac{W^{*NC}\left(p^{NC*}\right) - W^{NE}}{W^{NE}} \cdot 100,$$

$$INI^{ST} = \frac{W^{*ST}\left(p^{ST*}\right) - W^{NE}}{W^{NE}} \cdot 100.$$

All indices are relative measures, as absolute values are meaningless without any benchmark. Index *ICI* measures the potential gains from cooperation or what Barrett (1994) called the “need for cooperation”. Index *INI* measures the performance of stable coalitions. The paradox of cooperation comes in two versions. First, the *ICI* is small, and, hence, also the *INI* must be small, even though stable coalitions may be large. Second, the *ICI* may be large, but the *INI* is small because only small coalitions are stable. Hence, the “anti-paradox” would relate to large *ICI* and *INI*. That is, the potential gains from cooperation are large and these gains are reaped because large coalitions are stable.

Generally, and as our simulations will confirm, focusing only on the size of stable coalitions p^* may be misleading; also, the efficacy of cooperation needs to be evaluated.

We have conducted extensive simulations. The underlying simulation strategy is described in Appendix A.6. Those simulations are grouped into $\Psi < 0$ and $\Psi > 0$ of which Tables 2 and 3 show representative examples in that they capture all interesting features relevant for our discussion. Appendix A.6 explains further sensitivity analyses and refers to an Online Appendix where these additional results are available. Without any exception, all qualitative features displayed in Tables 2 and 3 are confirmed by our sensitivity analyses. In the tables, the most important columns are the one displaying the slope of the individual reaction function of non-signatories, $r_{NS}^{'}$, the size of stable coalitions, p^*, and the indexes *ICI* and *INI*, indicating the potential and actual gains from cooperation respectively. The other columns report about properties which have been discussed in previous sections.

Table 2, with $\Psi < 0$, confirms that for the ST-scenario, the steeper reaction functions in mitigation space are, the larger are stable coalitions. We recall that the sign of B_{aM} does not matter for this result. For the ST-scenario, large coalitions (including the grand coalition) may be stable, but the *ICI* and, hence, also the *INI* are small. Conversely, if the *ICI* is large, only small coalitions are stable and, hence, the *INI* is small. Importantly, this paradox of cooperation also holds for the NC-scenario because stable coalitions are always small, even if $B_{aM} > 0$ (instead of $B_{aM} < 0$) would be assumed. It is also evident that the ST-scenario only improves upon the NC-scenario by a margin if at all.

Table 3, with $\Psi > 0$, also confirms the paradox of cooperation for $B_{aM} < 0$. If the *ICI* is large, we have only $p^* = 3$ and if the *ICI* is small, we have $p^* = 3$ or $p^* = n$, with $n = 100$, but, of course, then also the *INI* is small, even for $p^{*}=n$.[18] Table 3 also confirms $p^{NC*}\geq p^{ST*}$, as stated in Proposition 3. The only case in which the paradox is not confirmed is if $p^{*}=n$ and $B_{aM}>0$. We interpret this as underlining the robustness of our conclusions because, as discussed in subsection 2.5, $B_{aM}>0$ is not very likely.

Taken together, the paradox of cooperation also holds true if adaptation is added to mitigation as a strategy to address climate change in a coalition formation game, regardless whether Stackelberg leadership of signatories is assumed. Even in those cases where large coalitions, including the grand coalition, are stable, stable coalitions improve only marginally upon no cooperation. This evaluation is missing in Bayramoglu et al. (2018). In other words, adaptation is not solving the paradox of cooperation.

6. Summary and Conclusion

In this paper, we considered the standard two-stage coalition formation game with symmetric players. We explored a mitigation-adaptation game under a Nash-Cournot scenario (NC-scenario) and a Stackelberg scenario (ST-scenario). In the first stage of the game, players choose whether to sign an agreement and be part of a climate coalition or to remain outside as a singleton. In the second stage, signatories choose their economic strategies (mitigation and adaptation) by maximizing their aggregate welfare, while non-signatories maximize their individual welfare. The sequence of these decisions differed between the NC- and the ST-scenario.

Our analysis combined features of two contributions. The first contribution by Barrett (1994), Diamantoudi and Sartzetakis (2006) and Rubio and Ulph (2006) who studied the effect of ST-scenario on the size of stable agreements in a pure mitigation (or emission) game; though in the absence of

[18] Whenever $p^{NC*}=p^{ST*}=3$, we should have $INI^{NC}<INI^{ST}$ according to Proposition 1b, but the differences are so small in our simulations that they hardly show up in Table 3, with values rounded to 2 digits. For $p^{NC*}=n>p^{ST*}=3$, the INI^{NC} is also only marginally larger than the INI^{ST} because the value of ICI is generally small.

adaptation. The second contribution by Bayramoglu et al. (2018) who studied the effect of adding adaptation to a mitigation coalition game under the NC-scenario.

We complemented these studies by considering Stackelberg leadership in a mitigation-adaptation game, which we viewed as a generalization of the pure mitigation game. This allowed us to address two research questions. 1) Does the ST-scenario improve over the NC-scenario? 2) Does the paradox of cooperation as established by Barrett (1994) and later reiterated by many others also hold if adaptation is included in the analysis?

We found that the ST-scenario leads to larger stable coalitions if reaction functions in mitigation space are downward sloping, i.e., mitigation levels in different countries are strategic substitutes. This happens because signatories reduce their mitigation efforts, forcing followers to mitigate more compared to the NC-scenario. Therefore, participation is more attractive in the ST- than in the NC-scenario. However, we found that whenever the difference in stable coalition sizes is large between the two scenarios, the potential gains from cooperation are small. Hence, the ST-scenario only marginally improves upon the NC-scenario. In contrast, if reaction functions in mitigation space are upward sloping, stable coalitions may be smaller in the ST- than in the NC-scenario, but in terms of global welfare the difference is again marginal. If large coalitions are stable, the gains from cooperation are small.

The results for the ST-scenario confirmed Barrett's paradox of cooperation: either coalitions are small or, if they are large, the potential gains from cooperation are small. Hence, the paradox extends to a game which includes adaptation. This is also true for the NC-scenario. Even though we confirm Bayramoglu et al. (2018) in that large coalitions can be stable in a mitigation-adaptation game, we qualify their positive conclusion because large stable coalitions only emerge if the gains from cooperation are small. Hence, the paradox of cooperation extends to a richer coalition game, which includes adaptation as an additional strategy to mitigation for the widespread assumption that mitigation and adaptation are substitutes.

For future research, two obvious extensions come to mind. Firstly, we assumed that adaptation is either chosen simultaneously with mitigation or after mitigation. In other words, we considered "reactive adaptation". However, in a dynamic game in which negotiations take place over some time and in which contracts are renegotiated, like for instance in Battaglini and Harstad (2016) and Harstad (2012), one can easily perceive that adaptation becomes "active" as considered for instance by Buob and Stephan (2011), Breton and Sbragia (2019) and Heuson et al. (2015). Secondly, we assumed symmetric players. In order to capture the current interesting discussion whether industrialized countries should support developing countries by providing adaptation because of their high vulnerability to climate change and their lack of adaptation capacity, the model would need to be extended to allow for asymmetry in terms of benefit and cost functions like this is considered for instance in Eyckmans et al. (2016), Lazkano et al. (2016) and Li and Rus (2018). However, as an anonymous reviewer pointed, the additional complexity would make it difficult to obtain analytical results.

References:

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

Battaglini, M. and B. Harstad (2016), Participation and Duration of Environmental Agreements. "Journal of Political Economy", vol. 124(1), pp. 160-204.

Baumol, W.J. and D.F. Bradford (1972), Detrimental Externalities and Non-Convexity of the Production Set. "Economica", vol. 39, pp. 160-176.

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

Breton, M. and L. Sbragia (2019), The Impact of Adaptation on the Stability of International Environmental Agreements. "Environmental and Resource Economics", vol. 74(2), pp. 697-725.

Brueckner, J.K. (2003), Strategic Interaction among Governments: an Overview of Empirical Studies. "International Regional Science Review", vol. 26(2), pp. 175-188.

Buob, S. and G. Stephan (2011), To Mitigate or to Adapt: How to Confront Global Climate Change. “European Journal of Political Economy”, vol. 27(1), pp. 1–16.

Carraro, C. and D. Siniscalco (1993), Strategies for the International Protection of the Environment. "Journal of Public Economics", vol. 52(3), pp. 309–328.

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.

Ebert, U. and H. Welsch (2012), Adaptation and Mitigation in Global Pollution Problems: Economic Impacts of Productivity, Sensitivity, and Adaptive Capacity. “Environmental and Resource Economics”, vol. 52, pp. 49–64.

Eichner, T. and R. Pethig (2015), Self-Enforcing Environmental Agreements and Trade in Fossil Energy Deposits. "Journal of Environmental Economics and Management", vol. 67(4), pp. 897-917.

Eichner, T. and R. Pethig (2017), Self-Enforcing International Environmental Agreements and Trade: Taxes versus Caps. "Oxford Economic Papers", vol. 85, pp. 1-20.

Eisenack, K. and L. Kähler (2016), Adaptation to Climate Change Can Support Unilateral Emission Reductions. "Oxford Economic Papers", vol. 68(1), pp. 258–278.

Endres, A. (1992), Strategic Behavior Under Tort Law. "International Reviw of Law and Economics", vol. 12, pp. 377-380.

Eyckmans, J., S. Fankhauser and S. Kverndokk (2016), Development Aid and Climate Finance. “Environmental and Resource Economics”, vol. 63 (2), pp. 429–450.

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

Fredriksson, P.G. and D. Millimet (2002a), Strategic Interaction and the Determination of Environmental Policy Across US States. "Journal of Urban Economics", vol. 51, pp. 101-122.

Fredriksson, P.G. and D. Millimet (2002b), Is There a "California Effect" in US Environmental Policymaking? "Regional Science and Urban Economics", vol. 32, pp. 737-764.

Gal-Or, E. (1985), First Mover and Second Mover Advantages. "International Economic Review", vol. 26(3), pp. 649-653.

Grubb, M.J., C. Hope and R. Fouquet (2002), Climatic Implications of the Kyoto Protocol: the

Contribution of International Spillover. "Climatic Change", vol. 54, pp.11-28.

Harstad, B. (2012), Climate Contracts: a Game of Emissions, Investments, Negotiations and Renegotiations. "Review of Econonomic Studies", vol. 79 (4), pp. 1527–1557.

Heuson, C., W. Peters, R. Schwarze and A.-K. Topp (2015), Investment and Adaptation as Commitment Devices. “Environmental and Resource Economics”, vol. 62, pp. 769-790.

Hoel, M. (1992), International Environment Conventions: The Case of Uniform Reductions of Emissions. “Environmental and Resource Economics”, vol. 2(2), pp. 141-159.

Independent Evaluation Group (IEG) (2013). Adapting to Climate Change: Assessing World Bank Group Experience: Phase III of the World Bank Group and Climate Change. Washington DC. http://ieg.worldbankgroup.org/sites/default/files/Data/Evaluation/files/cc3_full_eval.pdf

Ingham, A., J. Ma and A. Ulph (2013), Can Adaptation and Mitigation be Complements? "Climatic Change", vol. 120(1–2), pp. 39–53.

IPCC (2018), Summary for Policymakers. In: Global Warming of 1.5°C. An IPCC Special Report on the impacts of global warming of 1.5°C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change, sustainable development, and efforts to eradicate poverty [Masson-Delmotte, V., P. Zhai, H.-O. Pörtner, D. Roberts, J. Skea, P.R. Shukla, A. Pirani, W. Moufouma-Okia, C. Péan, R. Pidcock, S. Connors, J.B.R. Matthews, Y. Chen, X. Zhou, M.I. Gomis, E. Lonnoy, T. Maycock, M. Tignor, and T. Waterfield (eds.)]. World Meteorological Organization, Geneva, Switzerland, pp. 1-32.

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.

Lazkano, I., W. Marrouch and B. Nkuiya (2016), Adaptation to Climate Change: How Does Heterogeneity in Adaptation Costs Affect Climate Coalitions? “Environment and Development Economics”, vol. 21(06), pp. 812–838.

Li, H. and H. Rus (2018), Climate Change Adaptation and International Mitigation Agreements with Heterogeneous Countries. “Journal of the Association of Environmental and Resource Economists”, vol. 6(3), pp. 503-530.

Marrouch W. and A.R. Chaudhuri (2011), International Environmental Agreements in the Presence of Adaptation. "*FEEM Working Paper"* No. 35.2011.

Murdoch, J.C., T. Sandler and K. Sargent (1997), A Tale of Two Collectives: Sulphur versus Nitrogen Oxide Emission Reduction in Europe. "Economica", vol. 64, pp. 381-401.

Perkins, R. and E. Neumayer (2008), Fostering Environment-Efficiency Through Transnational Linkages? Trajectories of CO_2 and SO_2, 1980-2000. "Environment and Planning A", vol. 40(12), pp. 2970-2989.

Perkins, R. and E. Neumayer (2009), Transnational Linkages and the Spillover of Environment-Efficiency into Developing Countries. "Global Environmental Change", vol. 19(3), pp. 375-383.

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

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

Starrett, D.A. (1972), Fundamental Non-Convexities in the Theory of Externalities. "Journal of Economic Theory", vol.4, pp. 180-189.

UNFCCC (2014), Report of the Adaptation Committee to the Subsidiary Body for Scientific and Technological Advice. Forty-first session of COP20, Lima, Peru, FCCC/SB/2014/2.

UNFCCC (2016), Adaptation under the UNFCCC after the Paris Agreement. https://unfccc.int/news/adaptation-under-the-unfccc-after-the-paris-agreement

Weikard, H.-P. (2009), Cartel Stability under an Optimal Sharing Rule. “Manchester School”, vol. 77(5), pp. 575-593.

World Bank. (2010). Economics of Adaptation to Climate Change: Synthesis Report. Washington DC. http://documents.worldbank.org/curated/en/646291468171244256/pdf/702670ESW0P10800EACCSynthesisReport.pdf

Yohe, G. and K. Strzepek (2007), Adaptation and Mitigation as Complementary Tools for Reducing the Risk of Climate Impacts. "Mitigation and Adaptation Strategies for Global Change", vol. 12(5), pp. 727-739.

Zehaie, F. (2009) The Timing and Strategic Role of Self-Protection. "Environmental and Resource Economics", vol. 44(3), pp. 337–350.

Appendix

A.1 Technical Details around the Slopes of Reaction Functions in Mitigation Space

A.1.1 Second Order Conditions and Optimally Adapted Net Benefits

From the first order conditions as stated in Table 1, we derive the second order conditions in the NC-scenario:

Non-signatories: $B_{MM} - C_{mm} < 0$ *and signatories:* $p^2 \cdot B_{MM} - C_{mm} < 0$

with respect to mitigation and

Both: $B_{aa} - D_{aa} < 0$

with respect to adaptation. These conditions always hold due to the General Assumptions, as stated in section 2. Moreover, noting the interaction between mitigation and adaptation, substituting $a_i\,(m_i + M_{-i})$ into the payoff function, differentiating twice with respect to mitigation, noticing that $B_a = D_a$ from the first order conditions with respect to adaptation, we derive:[19]

Non-signatories: $\varPsi - C_{mm} < 0$ *and signatories:* $p^2 \cdot \varPsi - C_{mm} < 0$.

We note that these second order conditions are not automatically fulfilled if $\varPsi > 0$ but can generally be satisfied. We have omitted the arguments in these functions for convenience, but notice that $\varPsi = B_{MM} + \dfrac{(B_{aM})^2}{D_{aa} - B_{aa}}$ has the same value in both second order conditions because M^* and a_i^* will be the same for non-signatories and signatories. Moreover, if third derivatives are assumed to be zero, then $\varPsi$ and C_{mm} are constants, with $C_{mm}(m_{NS}^*) = C_{mm}(m_S^*)$. Consequently, in this case, $n^2 \cdot \varPsi - C_{mm} < 0$ is the most restrictive condition, which is the second order condition in the social optimum.

[19] We could also derive the Hessian matrix with the same result.

Accordingly, differentiating the optimally adapted net benefit function $B_i(M, a_i(M)) - D_i(a_i(M))$ twice, gives Ψ. Hence, the optimal adapted net benefit function is concave if $\Psi < 0$ and convex if $\Psi > 0$. Clearly, even if the optimally adapted net benefit function is convex, the second order conditions can be satisfied.

Finally, under the ST-scenario, everything is the same, except the second order condition of signatories with respect to mitigation, $p^2 \cdot \left(1 + R_{NS}^{'}\right) \cdot B_{MM} - C_{mm} < 0$, and those if $a_i(m_i + M_{-i})$ is considered, $p^2 \cdot \left(1 + R_{NS}^{'}\right) \cdot \Psi - C_{mm} < 0$, where for simplicity we have made use of the assumption that third derivatives are zero. Now it is clear that the first inequality is automatically satisfied due the General Assumptions. Moreover, one can show that if $n^2 \cdot \Psi - C_{mm} < 0$ holds, then also $p^2 \cdot \left(1 + R_{NS}^{'}\right) \cdot \Psi - C_{mm} < 0$ holds. If $R_{NS}^{'} < 0$, this is obvious. Hence, we assume $R_{NS}^{'} > 0$, insert $R_{NS}^{'}$ from (8) in the text into $p^2 \cdot \left(1 + R_{NS}^{'}\right)$, and show that this term is an increasing and convex function of p, using the condition $n^2 \cdot \Psi - C_{mm} < 0$. Thus, we insert the largest possible value for p which is $p = n - 1$ (recall if $p = n$, then there is no Stackelberg leadership) in $p^2 \cdot \left(1 + R_{NS}^{'}\right)$ and show that this is smaller than n^2. Hence, if the second order condition in the social optimum is satisfied, then also the second order condition of the Stackelberg leader is satisfied.

A.1.2 Existence and Uniqueness Condition for an Interior Equilibrium in the Second Stage

The procedure to derive sufficient conditions for the existence and uniqueness of mitigation and adaptation equilibria for every coalition of size p follows Bayramoglu et al. (2018). The procedure is based on the concept of replacement functions. 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 derived by summing over all replacement functions, which for symmetry is

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

If every replacement function is downward sloping over the entire mitigation space, the aggregate replacement function will be downward sloping as well (which is the vertical aggregation of individual replacement functions) and, hence, will intersect with the 45-degree line once. In other words, the level of M, which satisfies the equality above is the equilibrium M^*, which upon substitution into individual replacement functions gives m_S^* and m_{NS}^*. As we will see below, replacement functions are downward sloping (like reaction functions) if $\Psi < 0$. In the case of upward sloping replacement functions ($\Psi > 0$), a sufficient condition for uniqueness is that the aggregate replacement function has a slope of less than 1 over the entire domain of M such that it intersects with the 45-degree line and this happens only once. Note that if replacement functions are linear, which is the case if all third derivatives are zero, this sufficient condition is also a necessary condition. Finally, as reaction functions of adaptation as a function of total mitigation are continuous and single valued, also equilibrium adaptation levels will be unique. Below, we derive the sufficient conditions in the case of the ST-scenario, which are those in the NC-scenario as derived by Bayramoglu et al. (2018) if we set $R_{NS}^{'} = 0$.

The first order conditions of signatories with respect to mitigation and substituting $a_i\left(M\right)$, read:

$$p \cdot \left[B_M\left(M, a_i\left(M\right)\right) \cdot \left(1 + R_{NS}^{'}\right) \right] = C_m\left(m_S\left(M\right)\right)$$

using Table 1. Total differentiation with respect to M, and ignoring third derivatives for simplicity, gives the slope of the individual replacement function of signatories:

$$g_S^{'}\left(M\right) = \frac{p \cdot \left[\Psi \cdot \left(1 + R_{NS}^{'}\right) \right]}{C_{mm}\left(m_S\right)}.$$

For non-signatories, we find, using the first order conditions in Table 1:

$$B_M\left(M, a_i\left(M\right)\right) = C_m\left(m_{NS}\left(M\right)\right)$$

and, hence, we derive the slope of the individual replacement of non-signatories:

$$g_{NS}^{'}\left(M\right)=\frac{\varPsi}{C_{mm}\left(m_{NS}\right)}.$$

Accordingly, the slope of the aggregate replacement function is given by:

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

which is negative if $\varPsi<0$, but is positive if $\varPsi>0$. Hence, a sufficient condition for a unique interior equilibrium is $G^{'}(M)<1$ over the entire domain of M. In the NC-scenario, $R_{NS}^{'}=0$. If third derivative are zero, the largest value of $G^{'}(M)$ is if $p=n$ in which case the condition collapses to the second order condition in the social optimum, i.e., $n^{2}\cdot\Psi-C_{mm}<0$. Since the second order conditions as well as the condition for a unique interior equilibrium need to be satisfied for all values of p, $1\leq p\leq n$, $n^{2}\cdot\Psi-C_{mm}<0$ is the relevant condition, which we will use subsequently. For completeness let us point out that for any value $p\neq n$, the sufficient conditions for a unique interior equilibrium are more restrictive than the second order conditions. Finally, it is easily checked that if $n^{2}\cdot\varPsi-C_{mm}<0$ holds, also $G^{'}(M)<1$ under the ST-scenario, following the proof at the end of Appendix A.1.1.

A.1.3 Upper and Lower Bounds of the Slopes of Reaction Functions

Consider the slope of the reaction function in mitigation space of a single non-signatory, which is given by $r_{NS}^{'}\left(M_{-j}\right)=\frac{\varPsi}{C_{mm}\left(m_{NS}\right)-\varPsi}$. $C_{mm}\left(m_{NS}\right)-\varPsi>0$ by the condition $n^{2}\cdot\varPsi-C_{mm}<0$. If $\varPsi<0$, then $r_{NS}^{'}\left(M_{-j}\right)$ approaches -1 if C_{mm} approaches zero and approaches 0 if C_{mm} becomes very large. If $\varPsi>0$, then $r_{NS}^{'}\left(M_{-j}\right)$ increases in $\varPsi$, with the lower bound of $r_{NS}^{'}\left(M_{-j}\right)$ being 0 if $\varPsi$ approaches 0. However, the largest possible value of $\varPsi$ follows from $n^{2}\cdot\varPsi-C_{mm}<0$, which implies $\Psi < \frac{C_{mm}}{n^2}$. If we substitute $\Psi = \frac{C_{mm}}{n^2}$ into $r_{NS}^{'}\left(M_{-j}\right)$, we have: $r_{NS}^{'}\left(M_{-j}\right) = \frac{1}{(n-1)(n+1)}$ and this upper bound decreases with the number of players n. In our simulations, with $n = 100$, $\frac{1}{(n-1)(n+1)} = \frac{1}{9999} \approx 10^{-4}$.

A.2 Proof of Proposition 1

In a first step, we differentiate the left-hand side of signatories' first order conditions in mitigation space (5.a) under the ST-scenario with respect to M:

$$\frac{\partial\left[p\cdot\left(B_M\left(M,a_i\left(M\right)\right)\cdot\left(1+R_{NS}^{'}\right)\right)\right]}{\partial M} = p\cdot\left[\left[B_{MM} + B_{Ma}\cdot\frac{\partial a_i}{\partial M}\right]\cdot\left(1+R_{NS}^{'}\right)\right].$$

assuming second derivatives to be constant. Knowing that $\frac{\partial a_i}{\partial M} = \frac{B_{aM}}{D_{aa} - B_{aa}}$ and rearranging terms, we obtain:

$$\frac{\partial\left[p\cdot\left(B_M\left(M,a_i\left(M\right)\right)\cdot\left(1+R_{NS}^{'}\right)\right)\right]}{\partial M} = p\cdot\left[\Psi\cdot\left(1+R_{NS}^{'}\right)\right].$$

We notice that we would get the same for the NC-scenario (using (3.a) by setting $R_{NS}^{'} = 0$ above. Then, differentiating the benefit side of non-signatories' first order conditions (5.b), which is the same as (3.b), we obtain:

$$\frac{\partial\left[\left(B_M\left(M,a_i\left(M\right)\right)\right)\right]}{\partial M} = \left[B_{MM} + B_{Ma}\cdot\frac{\partial a_i}{\partial M}\right] = \Psi .$$

The signs of these derivatives depend on the sign of Ψ (as $1 + R_{NS}^{'} > 0$ is always true). Therefore, for both, signatories and non-signatories, the left-hand side of marginal benefits in their respective first order conditions will decrease (increase) in the level of total mitigation M if $\Psi < (>) 0$ under both scenarios.

1) Let us assume $\Psi < 0$. Hence, $R_{NS}^{'} < 0$. We want to show $M^{NC*}(p) > M^{ST*}(p)$ but assume the opposite: $M^{NC*}(p) \leq M^{ST*}(p)$.

From signatories' first order conditions under the NC-scenario (3.a) and under the ST-scenario (5.a), keeping in mind that for $\Psi < 0$, i.e., marginal benefits in the first order conditions decrease in total mitigation M, the following holds:

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

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

For non-signatories, using (3.b) or (5.b), which are identical, accordingly, we have:

$$C_m\left(m_{NS}^{ST*}\right) = B_M\left(M^{ST*}, a_i^{ST*}\left(M^{ST*}\right)\right) \leq B_M\left(M^{NC*}, a_i^{NC*}\left(M^{NC*}\right)\right) = C_m\left(m_{NS}^{NC*}\right).$$

It follows that $C_m\left(m_S^{ST*}\right) < C_m\left(m_S^{NC*}\right)$ and $C_m\left(m_{NS}^{ST*}\right) \leq C_m\left(m_{NS}^{NC*}\right)$ hold and, therefore, given the convexity of cost functions, $m_S^{ST*} < m_{NS}^{NC*}$ and $m_{NS}^{ST*} \leq m_{NS}^{NC*}$ must hold. These inequalities contradict the assumption $M^{NC*}(p) \leq M^{ST*}(p)$. Hence, $M^{NC*}(p) > M^{ST*}(p)$ must be true. Consequently, $m_{NS}^{NC*}(p) < m_{NS}^{ST*}(p)$ must hold from the first order conditions of non-signatories and, hence, for $M^{NC*}(p) > M^{ST*}(p)$, we must have $m_S^{NC*}(p) > m_S^{ST*}(p)$ for signatories.

Stackelberg leaders will be better off (or equal well off) than in the simultaneous game by axiomatic reasoning. For non-signatories, the variables that affect their welfare by going from the Nash-Cournot to the Stackelberg scenario are total mitigation (that also affects equilibrium adaptation levels) and individual mitigation. We know that mitigation costs will increase due to higher m_{NS}. In order to evaluate the overall effect, we totally differentiate a non-signatory's welfare function:

$$\Delta W_{NS} = \frac{\partial B\left(M^{NC}, a_i^{NC}\right)}{\partial M} \cdot \Delta M + \frac{\partial B\left(M^{NC}, a_i^{NC}\right)}{\partial a_i} \cdot \frac{\partial a_i}{\partial M} \cdot \Delta M - \frac{\partial C\left(m_{NS}^{NC}\right)}{\partial m_{NS}} \cdot \Delta m_{NS} - \frac{\partial D\left(a_i^{NC}\right)}{\partial M} \cdot \frac{\partial a_i}{\partial M} \cdot \Delta M$$

and, using the first order conditions in terms of adaptation, $B_a = D_a$, and dropping the arguments in the function above for convenience, we get:

$$\mathit{\Delta} W_{NS} = B_M \cdot \Delta M - C_m\left(m_{NS}\right) \cdot \Delta m_{NS} .$$

As we know from above that $\Delta M < 0$ and $\Delta m_{NS} > 0$, it follows that a non-signatory's welfare will drop when moving from the NC- to the ST-scenario. Therefore, pulling results together for $\mathit{\Psi} < 0$, $W_S^{NC*}\left(p\right) < W_S^{ST*}\left(p\right)$ and $W_{NS}^{NC*}\left(p\right) > W_{NS}^{ST*}\left(p\right)$ hold, though nothing can be said about aggregate welfare $W^*\left(p\right)$ at a general level. Noting that $W_{NS}^{NC*}\left(p\right) > W_{NS}^{ST*}\left(p\right)$ holds for every p, $1 < p < n$, we also have $W_{NS}^{NC*}\left(p-1\right) > W_{NS}^{ST*}\left(p-1\right)$. Considering internal stability, $W_S^*\left(p\right) \geq W_{NS}^*\left(p-1\right)$, we notice that the left-hand side term is larger and the right-hand side term smaller under the ST-scenario than under the NC-scenario. Hence, $p^{ST*} \geq p^{NC*}$ follows. This conclusion is still true if we consider the boundary values of p, namely $p = n$, in which case $W_S^*\left(p\right)$ is the same under both scenarios but $W_{NS}^*\left(p-1\right)$ is lower under the ST- than NC-scenario, and $p = 2$, in which case $W_{NS}^*\left(p-1\right)$ is the same under both scenarios according to our assumption, but $W_S^*\left(p\right)$ is larger under the ST- than NC-scenario. Finally, under the ST-scenario, strict superadditivity always holds (see Proposition 2). For the move from $p-1=1$ to $p=2$, this implies $2 \cdot W_S^*(2) > 2 \cdot W_{NS}^*(1)$ or $W_S^*(2) > W_{NS}^*(1)$ and the condition for internal stability requires $W_S^*(2) \geq W_{NS}^*(1)$. Thus, $p^{ST*} \geq 2$.

2) We now consider $\mathit{\Psi} > 0$. We want to show $M^{NC*}\left(p\right) < M^{ST*}\left(p\right)$.

From the first order conditions of signatories (3.a) and (5.a), it is clear that $M^{NC*}\left(p\right) = M^{ST*}\left(p\right)$ is not possible. Due to upward-sloping mitigation reaction functions, we need to consider two possibilities:

$M^{NC*}(p) < M^{ST*}(p)$, which would be compatible only with $m_S^{NC*}(p) < m_S^{ST*}(p)$ and $m_{NS}^{NC*}(p) < m_{NS}^{ST*}(p)$;

$M^{NC*}(p) > M^{ST*}(p)$, which would be compatible only with $m_S^{NC*}(p) > m_S^{ST*}(p)$ and $m_{NS}^{NC*}(p) > m_{NS}^{ST*}(p)$.

We note that, axiomatically, the Stackelberg leader will receive a higher (or equal) payoff compared to the NC-scenario. To see how signatories' welfare will change when moving from the NC- to the ST-scenario, we total differentiate welfare function (1). The result would be the same for non-signatories, except for individual mitigation levels (as done below). We have:

$$\mathit{\Delta W}_S = \frac{\partial B\left(M^{NC}, a_i^{NC}\right)}{\partial M} \cdot \Delta M + \frac{\partial B\left(M^{NC}, a_i^{NC}\right)}{\partial a_i} \cdot \frac{\partial a_i}{\partial M} \cdot \Delta M - \frac{\partial C\left(m_S^{NC}\right)}{\partial m_S} \cdot \Delta m_S - \frac{\partial D\left(a_i^{NC}\right)}{\partial M} \cdot \frac{\partial a_i}{\partial M} \cdot \Delta M$$

and, using the information $B_a = D_a$ from the first order conditions with respect to adaptation, we get:

$$\mathit{\Delta W}_S = B_M \cdot \Delta M - C_m\left(m_S\right) \cdot \Delta m_S .$$

From the first order conditions of signatories under the NC-scenario (3.a) in Table 1, we know that $p \cdot B_M = C_m\left(m_S\right)$. Hence, $\mathit{\Delta W}_S = B_M \cdot \Delta M - B_M p \cdot \Delta m_S = B_M\left(\Delta M - p \cdot \Delta m_S\right)$. In the case of $M^{NC*}(p) < M^{ST*}(p)$, $\Delta M > 0$, $\Delta m_S > 0$ and $\Delta m_{NS} > 0$ with $\Delta M > p \cdot \Delta m_s$ and, hence, $\mathit{\Delta W}_S > 0$ follows. In the case of $M^{NC*}(p) > M^{ST*}(p)$, $\Delta M < 0$, $\Delta m_S < 0$ and $\Delta m_{NS} < 0$ with $\left|\Delta M\right| > \left|p \cdot \Delta m_s\right|$ and, hence, $\mathit{\Delta W}_S < 0$, which must be wrong by the axiomatic reasoning above.

Therefore, for $\Psi > 0$, we will have: $M^{NC*}(p) < M^{ST*}(p)$, $m_S^{NC*}(p) < m_S^{ST*}(p)$ and $m_{NS}^{NC*}(p) < m_{NS}^{ST*}(p)$.

For non-signatories, we have:

$$\Delta W_{NS} = B_M \cdot \Delta M - C_m\left(m_{NS}\right) \cdot \Delta m_{NS} .$$

From the first order conditions of non-signatories under the NC-scenario (3.b) in Table 1, we know that $B_M = C_m$. We also know that because of upward sloping mitigation reaction functions $|\Delta M| > |\Delta m_{NS}|$ holds and, hence, $|B_M \cdot \Delta M| > |C_m(m_{NS}) \cdot \Delta m_{NS}|$. Hence, taken together, $W_S^{NC*}(p) < W_S^{ST*}(p)$ and $W_{NS}^{NC*}(p) < W_{NS}^{ST*}(p)$. Hence, $W^{NC*}(p) < W^{ST*}(p)$ if $\Psi > 0$.

Now, we need to show $m_S^{ST*}(p) - m_S^{NC*}(p) > m_{NS}^{ST*}(p) - m_{NS}^{NC*}(p)$, which implies $W_S^{ST*}(p) - W_S^{NC*}(p) < W_{NS}^{ST*}(p) - W_{NS}^{NC*}(p)$. Looking at signatories' and non-signatories' welfare functions, we can rewrite those as follows: $W_{NS}^{NC*} = W_S^{NC*} + \left(C\left(m_S^{NC*}\right) - C\left(m_{NS}^{NC*}\right)\right)$ and $W_{NS}^{ST*} = W_S^{ST*} + \left(C\left(m_S^{ST*}\right) - C\left(m_{NS}^{ST*}\right)\right)$. Using this, $W_S^{ST*}(p) - W_S^{NC*}(p) < W_{NS}^{ST*}(p) - W_{NS}^{NC*}(p)$ translates into $C\left(m_S^{ST*}\right) - C\left(m_S^{NC*}\right) > C\left(m_{NS}^{ST*}\right) - C\left(m_{NS}^{NC*}\right)$. This will be true provided $m_S^{ST*}(p) - m_S^{NC*}(p) > m_{NS}^{ST*}(p) - m_{NS}^{NC*}(p)$ holds, which we need to prove.

We know from above that for $\Psi > 0$, $M^{ST*} > M^{NC*}$ and $B_M\left(M^{ST*}, a_i^{ST*}\right) > B_M\left(M^{NC*}, a_i^{NC*}\right)$. Hence, from the FOCs of signatories and non-signatories, we have:

$$C_m\left(m_S^{ST*}\right) = C_m\left(m_S^{NC*}\right) + p \cdot \left[B_M\left(M^{ST*}, a_i^{ST*}\right) - B_M\left(M^{NC*}, a_i^{NC*}\right) + B_M\left(M^{ST*}, a_i^{ST*}\right) \cdot R_{NS}^{'}\right] \text{ and}$$

$$C_m\left(m_{NS}^{ST*}\right) = C_m\left(m_{NS}^{NC*}\right) + B_M\left(M^{ST*}, a_i^{ST*}\right) - B_M\left(M^{NC*}, a_i^{NC*}\right).$$

Moving the cost terms to the left-hand side, we have:

$$C_m\left(m_S^{ST*}\right) - C_m\left(m_S^{NC*}\right) = p \cdot \left[B_M\left(M^{ST*}, a_i^{ST*}\right) - B_M\left(M^{NC*}, a_i^{NC*}\right) + B_M\left(M^{ST*}, a_i^{ST*}\right) \cdot R_{NS}^{'}\right] \quad \text{(A.1)}$$

$$C_m\left(m_{NS}^{ST*}\right) - C_m\left(m_{NS}^{NC*}\right) = B_M\left(M^{ST*}, a_i^{ST*}\right) - B_M\left(M^{NC*}, a_i^{NC*}\right). \quad \text{(A.2)}$$

We know that $C_m\left(m_S^{ST*}\right) - C_m\left(m_S^{NC*}\right) > 0$ as $m_S^{ST*} > m_S^{NC*}$ and $C_m\left(m_{NS}^{ST*}\right) - C_m\left(m_{NS}^{NC*}\right) > 0$ as $m_{NS}^{ST*} > m_{NS}^{NC*}$ if $\Psi > 0$.

Substituting (A.2) into (A.1), we have:

$$C_m\left(m_S{}^{ST*}\right)-C_m\left(m_S{}^{NC*}\right)=p\cdot\left[C_m\left(m_{NS}{}^{ST*}\right)-C_m\left(m_{NS}{}^{NC*}\right)+B_M\left(M^{ST*},a_i^{ST*}\right)\cdot R_{NS}^{'}\right]. \qquad \text{(A.3)}$$

Assuming second derivatives to be constant, the differences in marginal costs can be rewritten as:

$C_m\left(m_S{}^{ST*}\right)-C_m\left(m_S{}^{NC*}\right)=C_{mm}\cdot\left[m_S{}^{ST*}-m_S{}^{NC*}\right]$ and

$$C_m\left(m_{NS}{}^{ST*}\right)-C_m\left(m_{NS}{}^{NC*}\right)=C_{mm}\cdot\left[m_{NS}{}^{ST*}-m_{NS}{}^{NC*}\right].$$

Substituting into (A.3), we obtain:

$$C_{mm}\cdot\left[m_S{}^{ST*}-m_S{}^{NC*}\right]=p\cdot\left[C_{mm}\cdot\left[m_{NS}{}^{ST*}-m_{NS}{}^{NC*}\right]+B_M\left(M^{ST*},a_i^{ST*}\right)\cdot R_{NS}^{'}\right].$$

Finally, dividing through by C_{mm}, we have:

$$m_S{}^{ST*}-m_S{}^{NC*}=p\cdot\left[m_{NS}{}^{ST*}-m_{NS}{}^{NC*}+\frac{B_M\left(M^{ST*},a_i^{ST*}\right)\cdot R_{NS}^{'}}{C_{mm}}\right]$$

From this inequality, we can conclude $m_S^{ST*}\left(p\right)-m_S^{NC*}\left(p\right)>m_{NS}^{ST*}\left(p\right)-m_{NS}^{NC*}\left(p\right)$ as $R_{NS}^{'}>0$ if $\Psi>0$ and, hence, $W_S^{ST*}\left(p\right)-W_S^{NC*}\left(p\right)<W_{NS}^{ST*}\left(p\right)-W_{NS}^{NC*}\left(p\right)$.

Finally, $p^{NC*}\geq 2$ and $p^{ST*}\geq 2$ follows from the fact that for $\Psi>0$ the game is a superadditive coalition game for both scenarios according to Proposition 2. Then, we apply the same proof as outlined above.

Remark: Note that for $\Omega^{ST}(p):=W_S^{ST*}(p)-W_{NS}^{ST*}(p-1)$ and $\Omega^{NC}(p):=W_S^{NC*}(p)-W_{NS}^{NC*}(p-1)$, $\Omega^{ST}(n)<\Omega^{NC}(n)$ because $W_S^{ST*}(n)=W_S^{NC*}(n)$ and $W_{NS}^{ST*}(n-1)>W_{NS}^{NC*}(n-1)$, and $\Omega^{ST}(2)>\Omega^{NC}(2)$ because $W_S^{ST*}(2)>W_S^{NC*}(2)$ and $W_{NS}^{ST*}(1)=W_{NS}^{NC*}(1)$. Thus, stability functions $\Omega\ (p)$ under the two scenarios cross each other at least once for $\Psi>0$, which makes general predictions difficult. As will be apparent from Appendix A.5 for the specific payoff function (10a) and (10b), the curvature of the stability functions $\Omega^{ST}(p)$ and $\Omega^{NC}(p)$ under the two scenarios may be quite complex, with fluctuating upward and downward sloping as well as convex and concave segments in different intervals of p .

A.3 Proof of Proposition 2

Mitigation Cohesiveness (MCOH) and the Change of Signatories' and Non-signatories Equilibrium Mitigation Levels with Membership

The difference $M^*(p)-M^*(p-1)$ can also be investigated by considering $\frac{\partial M^*}{\partial p}$, treating p as a continuous variable. Bayramoglu et al. (2018) have derived $\frac{\partial M^*}{\partial p}$ in the NC-scenario. Following their approach, only minor modifications for the ST-scenario are necessary. By setting $R_{NS}^{'}=0$, we retrieve the conditions in the NC-scenario. Total differentiation of the first order conditions of signatories and non-signatories, as provided in Table 1, assuming second derivatives to be constant, delivers:

$$\frac{\partial m_S^*}{\partial p}=\frac{p\cdot\Psi\cdot\frac{\partial M^*}{\partial p}\cdot\left(1+R_{NS}^{'}\right)}{C_{mm}\left(m_S^*\right)}+\frac{B_M\cdot\left(1+R_{NS}^{'}\right)}{C_{mm}\left(m_S^*\right)}$$

$$\frac{\partial m_{NS}^*}{\partial p}=\frac{\Psi\cdot\frac{\partial M^*}{\partial p}}{C_{mm}\left(m_{NS}^*\right)}$$

where $\frac{B_M\cdot\left(1+R_{NS}^{'}\right)}{C_{mm}\left(m_S^*\right)}>0$. We have: $\frac{\partial M^*}{\partial p}=m_S^*+p\cdot\frac{\partial m_S^*}{\partial p}-m_{NS}^*+\left(n-p\right)\cdot\frac{\partial m_{NS}^*}{\partial p}$. Substituting $\frac{\partial m_S^*}{\partial p}$ and $\frac{\partial m_{NS}^*}{\partial p}$ from above and rearranging terms, we obtain:

$$\frac{\partial M^*}{\partial p}=\frac{m_S^*-m_{NS}^*+\frac{p\cdot B_M\cdot\left(1+R_{NS}^{'}\right)}{C_{mm}\left(m_S^*\right)}}{1-\Psi\cdot\left[\frac{p^2\cdot\left(1+R_{NS}^{'}\right)}{C_{mm}\left(m_S^*\right)}+\frac{\left(n-p\right)}{C_{mm}\left(m_{NS}^*\right)}\right]}$$

The term $\frac{p \cdot B_M \cdot \left(1+R_{NS}^{'}\right)}{C_{mm}\left(m_S^*\right)}$ is always positive and the denominator is always positive by the sufficient conditions for the existence of a unique interior equilibrium, as derived in Appendix A.1. Hence, if $m_S^* - m_{NS}^* > 0$, we can conclude $\frac{\partial M^*}{\partial p} > 0$. We know that $m_S^* - m_{NS}^* > 0$ if $\Psi > 0$ in both scenarios, in which case we can also conclude $\frac{\partial m_S^*}{\partial p} > 0$ and $\frac{\partial m_{NS}^*}{\partial p} > 0$ from above. If $\Psi < 0$, in the NC-scenario, we also have $m_S^* - m_{NS}^* > 0$. Hence, $\frac{\partial M^*}{\partial p} > 0$ and $\frac{\partial m_{NS}^*}{\partial p} < 0$ can be concluded, but nothing can be concluded about $\frac{\partial m_S^*}{\partial p}$, as the first term is negative and the second positive in the derivative above. If $\Psi < 0$, in the ST-scenario, $m_S^* - m_{NS}^* < 0$ (which typically happens for small values of p) and $m_S^* - m_{NS}^* > 0$ (which typically happens for sufficiently large values of p) are possible. Hence, nothing can be generally concluded about the signs of $\frac{\partial M^*}{\partial p}$, $\frac{\partial m_{NS}^*}{\partial p}$ and $\frac{\partial m_S^*}{\partial p}$.

Positive Externality Property (PEP) and Positive Internalisation Property (PIP)

Let us first consider PEP, again, treating p as a continuous variable. In the context of the NC-scenario, see Bayramoglu et al. (2018). In the ST-scenario, we derive exactly the same condition for non-signatories:

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

noting that $B_M > 0$ from the General Assumptions and $\left(1 - \frac{\Psi}{C_{mm}\left(m_{NS}^{ST*}\right)}\right) > 0$ from the sufficient condition of the existence of a uniqueness interior equilibrium, as stated in Appendix A.1. Therefore, $\frac{\partial W_{NS}^{ST*}}{\partial p}$ depends on the sign of $\frac{\partial M^{ST*}}{\partial p}$. Whereas $\frac{\partial M^{NC*}}{\partial p} > 0$ always holds in the NC-scenario, and

$\dfrac{\partial M^{ST*}}{\partial p} > 0$ in the ST-scenario if $\Psi > 0$, as we know from above, we also know that in the ST-scenario $\dfrac{\partial M^{ST*}}{\partial p} < 0$ is possible if $\Psi < 0$ in which case non-signatories do not enjoy a positive but suffer from a negative externality if the coalition is expanded.

Let us now consider PIP, again, treating p as a continuous variable. In the NC-scenario, we have $\dfrac{\partial W_S^{NC*}}{\partial p} = B_M \cdot \left[\dfrac{\partial M}{\partial p}^{NC*} - p \cdot \dfrac{\partial m_S^{NC*}}{\partial p}\right]$. Using $\dfrac{\partial M^{NC*}}{\partial p} = m_S^{NC*} + p \cdot \dfrac{\partial m_S^{NC*}}{\partial p} - m_{NS}^{NC*} + (n-p) \cdot \dfrac{\partial m_{NS}^{NC*}}{\partial p}$, we get:

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

In the NC-scenario, we always have: $m_S^{NC*} - m_{NS}^{NC*} > 0$, while $\dfrac{\partial m_{NS}^{NC*}}{\partial p} > (<) 0$ if $\Psi > (<) 0$. Hence, $\dfrac{\partial W_S^{NC*}}{\partial p} > 0$ if $\Psi > 0$. If $\Psi < 0$, we have $m_S^{NC*} - m_{NS}^{NC*} > 0$ and $\dfrac{\partial m_{NS}^{NC*}}{\partial p} < 0$. In this case, PIP may fail for small p because $m_S^{NC*} - m_{NS}^{NC*}$ is small and because $(n-p)\dfrac{\partial m_{NS}^{NC*}}{\partial p}$ is large in absolute terms.

In the ST-scenario, we have: $\dfrac{\partial W_S^{ST*}}{\partial p} = B_M \cdot \left[\dfrac{\partial M}{\partial p}^{ST*} - p \cdot \left[1 + R_{NS}^{'}\right] \cdot \dfrac{\partial m_S^{ST*}}{\partial p}\right]$ or

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

The sign of PIP depends on the sign of the term in brackets. Substituting $\dfrac{\partial m_{NS}^{ST*}}{\partial p}$ and $\dfrac{\partial m_S^{ST*}}{\partial p}$ from above and using $R_{NS}^{'} = \dfrac{(n-p) \cdot \Psi}{C_{mm}\left(m_{NS}^{ST}\right) - (n-p) \cdot \Psi}$, we have:

$$\frac{\partial W_S^{ST*}}{\partial p} = B_M\left[\frac{\left(m_S^{ST*}-m_{NS}^{ST*}\right)\cdot C_{mm}\left(m_S^{ST*}\right)}{C_{mm}\left(m_{NS}^{ST*}\right)-\left(n-p\right)\cdot\Psi}\right] \text{ with } \Psi = B_{MM}+\frac{\left(B_{aM}\right)^2}{D_{aa}-B_{aa}}$$

The denominator is positive for the existence-uniqueness condition (see Appendix A.1.2). In the nominator, $\left(B_{aa}-D_{aa}\right)\cdot C_{mm}<0$ from the General Assumptions. Hence, $\frac{\partial W_S^{ST*}}{\partial p}>0$ if $m_{NS}^{ST*}-m_S^{ST*}<0$ (which is always true if $\Psi>0$) and the reverse is true, i.e., $\frac{\partial W_S^{ST*}}{\partial p}<0,$ if $m_{NS}^{ST*}-m_S^{ST*}>0$ (which could happen if $\Psi<0$).

Superadditivity (SAD)

We need to show: $p\cdot W_S^*\left(p\right)\geq(>)\left[p-1\right]\cdot W_S^*\left(p-1\right)+W_{NS}^*\left(p-1\right)$ for all p, $2\leq p\leq n$. For the NC-scenario Bayramoglu et al. (2018) established that a sufficient condition for SAD to hold are (weakly) upward sloping reaction functions, i.e., $\Psi\geq 0$. For the ST-scenario, we notice that SAD must hold by axiomatic reasoning. Step 1: Any move from $p-1$ to p implies one more signatory. Keeping total mitigation of the p signatories at the same level than at $p-1$ $\left(p\cdot\tilde{m}_S(p)=\left[p-1\right]m_S^*(p-1)+m_{NS}^*(p-1)\right)$, total mitigation cost will have decreased among the p signatories as the first order conditions of mitigation imply cost-effectiveness among signatories. The $n-p$ non-signatories will not have changed their strategies in Step 1. Step 2: The p Stackelberg leaders choose their equilibrium strategies by maximizing their aggregate payoff, taking the best-response of non-signatories into account. If they choose different strategies in step 2 compared to step 1 $(m_S^*(p)\neq\tilde{m}_S(p))$, the aggregate welfare of the p signatories must have further increased. For the final move from $p-1=n-1$ to $p=n$, when there are no outsiders left after the move, the SAD-condition is equal to welfare cohesiveness (WCOH) and WCOH for this last move does generally hold because total welfare in the grand coalition is strictly larger than in any other coalition in an externality game by axiomatic reasoning.

Welfare Cohesiveness (WCOH)

If a game is superadditive and exhibits a positive externality throughout, this is sufficient that WCOH holds. Both conditions hold in both scenarios for $\Psi > 0$. In order to prove that WCOH may fail for $\Psi < 0$, we note that under the NC-scenario SAD may fail and under the ST-scenario PEP may fail, and refer to Table 2 for examples.

A.4 Equilibrium Mitigation and Adaptation for Payoff Function (7.a) and (7.b)

Due to space limitations, we only provide the central equations here, all details are provided in an Online Appendix 1 (NC-scenario) and an Online Appendix 2 (ST-scenario). Considering payoff function (10.a) for which $B_{aM} < 0$, we have: $B_M = b - g \cdot M - f \cdot a_i$, $B_{MM} = -g < 0,$ $B_{Ma} = -f < 0$, $B_a = \beta - f \cdot M$, $B_{aa} = 0$, $C_m = c \cdot m_i$, $C_{mm} = c,$ $D_a = d \cdot a_i$, $D_{aa} = d$ and $\Psi = -g + \frac{(-f)^2}{d} = \frac{f^2 - g \cdot d}{d}$. The sign of Ψ depends on the sign of $f^2 - g \cdot d$. From the sufficient condition for the existence of a unique interior equilibrium under the NC-scenario $\Psi \cdot \left[\frac{p^2}{C_{mm}(m_S)} + \frac{(n-p)}{C_{mm}(m_{NS})} \right] < 1$ (see Appendix A.1), noticing that $C_{mm}(m_S) = C_{mm}(m_{NS}) = c$ as well as Ψ are constants, the left-hand side of this inequality increases in p. Hence, using $p = n$, we derive for payoff function (10a) $c \cdot d - n^2 \cdot \left(f^2 - g \cdot d \right) > 0$ for this condition. We notice that this condition is not binding if $f^2 - g \cdot d < 0$, i.e., $\Psi < 0$, as expected. Moreover, due to linear replacement functions, if this condition is binding (i.e., $\Psi > 0$), it is also a necessary condition. From Appendix A.1 we may recall that this condition is also sufficient to guarantee a unique interior equilibrium under the ST-scenario.

For the slopes of the reaction functions, we derive:

$$r_S^{'}\left(M_{-i}\right)=\frac{p\cdot\left(f^2-d\cdot g\right)}{c\cdot d-p\cdot\left(f^2-d\cdot g\right)},\ R_S^{'}\left(M_{NS}\right)=\frac{p^2\cdot\left(f^2-d\cdot g\right)}{c\cdot d-p^2\cdot\left(f^2-d\cdot g\right)},\ r_{NS}^{'}\left(M_{-j}\right)=\frac{f^2-d\cdot g}{c\cdot d-\left(f^2-d\cdot g\right)},$$

$$R_{NS}^{'}\left(M_S\right)=\frac{\left(n-p\right)\cdot\left(f^2-d\cdot g\right)}{c\cdot d-\left(n-p\right)\cdot\left(f^2-d\cdot g\right)}\text{ and }h^{'}\left(M\right)=\frac{-f}{d}.$$

where $r^{'}$ denotes the slopes of individual and $R^{'}$ the slopes of the aggregate reaction functions. For the NC-scenario, we have:

$$m_S^{NC*}(p)=\frac{p\cdot\left(b\cdot d-\beta\cdot f\right)}{c\cdot d-\left(p^2+n-p\right)\cdot\left(f^2-d\cdot g\right)},\ m_{NS}^{NC*}(p)=\frac{m_S^{NC*}(p)}{p}\text{ and}$$

$$a_i^{NC*}(p)=\frac{\beta\cdot c-\left(n-p+p^2\right)\cdot\left(b\cdot f-g\cdot\beta\right)}{c\cdot d-\left(n-p+p^2\right)\cdot\left(f^2-d\cdot g\right)}.$$

In the ST-scenario, we find:

$$m_S^{ST*}(p)=\frac{c\cdot p\cdot d\cdot(b\cdot d-\beta\cdot f)}{Z},\ m_{NS}^{ST*}(p)=\frac{\left(c\cdot d-(n-p)\cdot\left(f^2-g\cdot d\right)\right)\cdot(b\cdot d-\beta\cdot f)}{Z}\text{ and}$$

$$a_i^{ST*}(p)=\frac{(b\cdot f-g\cdot\beta)\cdot(f^2-d\cdot g)\cdot(n-p)^2+c\cdot\beta\cdot(c\cdot d+\left(n-p\right)f^2)-X}{Z}$$

with $Z:=(f^2-d\cdot g)\cdot\left(-c\cdot d\cdot(p^2+2n-2p)\right)+(f^2-d\cdot g)^2\cdot(n-p)^2+c^2\cdot d^2$ and

$X:=b\cdot f\cdot c\cdot d\cdot(n-p+p^2)+g\cdot\beta\cdot c\cdot d\cdot(2(n-p)+p^2)$.

We impose the following five conditions on parameters:

$C1^{ST}=C1^{NC}:\ b-g\cdot M-f\cdot a>0$,

$C2^{ST}=C2^{NC}:\ \beta-f\cdot M>0$,

$C3^{ST}=C3^{NC}:\ c\cdot d-n^2\cdot\left(f^2-g\cdot d\right)>0$,

$C4^{ST}=C4^{NC}:\ b\cdot d-\beta\cdot f>0$,

$C5^{NC}:\ \beta\cdot c - n^2\cdot\left(b\cdot f - g\cdot\beta\right) > 0$,

$$C5^{ST}:\ (b\cdot f-\beta\cdot g)\cdot(f^2-d\cdot g)\cdot(n-p)^2+c\cdot\beta\cdot(f^2\cdot p-f^2\cdot n+c\cdot d)-b\cdot f\cdot c\cdot d\cdot(p^2+n-p)+ \\ g\cdot\beta\cdot c\cdot d\cdot(p^2+2n-2p)>0\,.$$

Conditions $C1$ and $C2$ are required for the General Assumptions to hold (i.e., $B_M > 0$ and $B_a > 0$; see Section 2); $C3$ is the condition for the existence of a unique interior equilibrium as discussed already above; $C4$ and $C5$ are the respective additional conditions that ensure that equilibrium mitigation and adaptation levels are positive. $C4$ ensures that the numerators of equilibrium mitigation levels are positive. (The remaining term in the numerator of $m_{NS}^{ST*}(p)$ is positive due to the condition for the existence of a unique interior equilibrium, i.e., $C3^{ST} = C3^{NC}$.) $C5$ ensures that the numerators of equilibrium adaptation levels are positive. Finally, we note that the term Z, the denominator of equilibrium mitigation and adaptation levels, is always positive. The second term in Z, $(f^2-d\cdot g)^2\cdot(n-p)^2$, is always positive. Hence, we have to sign $c^2\cdot d^2$ $-\left(c\cdot d\cdot(p^2+2n-2p)\cdot(f^2-d\cdot g)\right)$. Dividing by $c\cdot d$, we obtain $c\cdot d$ $-(p^2+2n-2p)\cdot(f^2-d\cdot g)$, which takes on its lowest value for $p=n$. Replacing $p=n$, we obtain $C3^{ST} = C3^{NC}$.

Substituting the highest possible equilibrium mitigation and adaptation levels in $C1$ and $C2$, it turns out that these two conditions are captured by the non-negativity conditions $C4$ and $C5$. Therefore, for both scenarios, only condition $C3$ to $C5$ are relevant, with $C3$ being only relevant if $f^2 - g\cdot d > 0$, i.e., if $\Psi > 0$.

Moving now to the case of $B_{aM} > 0$ i.e., considering payoff function (10.b), it turns out that some of the conditions above can be dropped (e.g., $C4$ because f is now $-f$ and $b\cdot d+\beta\cdot f>0$ is always true) and no additional conditions need to be imposed.

A.5 Proof of Proposition 3

Due to space limitations, we only provide the central idea of the proof; all details are provided in an Online Appendix 3 (NC-scenario) and an Online Appendix 4 (ST-scenario).

The result for $\Psi < 0$ and the NC-scenario is proved in Bayramoglu et al. (2018). For the ST-scenario, $p^* \geq 2$ follows from the fact that $p = 2$ is internally stable by the property superadditivity. Thus, if $p = 2$ is not externally stable, some larger coalition will be stable. The fact that any coalition between $p = 2$ and $p = n$ follows from our simulations on which we report in the paper.

Consider now $\Psi > 0$.

Bayramoglu et al. (2018) derive the sign of $\Omega^{NC}(p) := W_S^{NC*}(p) - W_{NS}^{NC*}(p-1)$ for the NC-scenario:

$$sign\left[\Omega^{NC}(p,n)\right] = -\left[\Phi_1(p,n)\cdot\Psi^2 - c\cdot\Phi_2(p,n)\cdot\Psi + (p-3)\cdot c^2\right]$$

with:

$\Phi_1(p,n) = p^5 - 5p^4 + 2np^3 + 7p^3 - 4np^2 + n^2p - 3p^2 - 2np + n^2$,

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

$\Psi = \dfrac{f^2 - gd}{d}$, with Ψ derived in Appendix A.4. If $sign\left[\Omega^{NC}(p,n)\right] \geq 0$, internal stability holds. We assume generally, $n \geq 7$. We note that $p = 1$ and $p = 2$ are internally stable, but not externally stable. Hence, we focus on $p \geq 3$.

It can be shown that $\Phi_1(p,n) > 0$ for any p and $\Phi_2(p,n) > 0$ for $p \geq 3$.

For $p = 3$, $sign\left[\Omega^{NC}(3,n)\right] = -\left[\Phi_1(3,n)\cdot\Psi^2 - c\cdot\Phi_2(3,n)\cdot\Psi\right]$ with $\Phi_1(3,n) = 4n^2 + 12n$ and $\Phi_2(3,n) = 4n - 4$. Solving $sign\left[\Omega^{NC}(3,n)\right] \geq 0$ gives $\Psi \leq \dfrac{c\cdot\Phi_2}{\Phi_1} = c\dfrac{4n-4}{4n^2+12n}$. By the sufficient conditions for the existence of a unique equilibrium, $\Psi < \frac{c}{n^2}$, this condition always holds because $\frac{4n-4}{4n^2+12n} \geq \frac{1}{n^2}$ for $n \geq 3$. Thus, $sign\left[\Omega^{NC}(3,n)\right] > 0$.

For $p > 3$, $sign\left[\Omega^{NC}(p,n)\right] = -\left[\Phi_1(p,n)\cdot\Psi^2 - c\cdot\Phi_2(p,n)\cdot\Psi\right] \geq 0$ for $\Psi \in [\underline{\Psi}^{NC};\bar{\Psi}^{NC}]$ with $\underline{\Psi}^{NC} = -\frac{(-\Phi_2 + \sqrt{-4\Phi_1\cdot p + \Phi_2^2 + 12\Phi_1})\cdot c}{2\Phi_1}$ and $\bar{\Psi}^{NC} = -\frac{(\Phi_2 + \sqrt{-4\Phi_1\cdot p + \Phi_2^2 + 12\Phi_1})\cdot c}{2\Phi_1}$.

For $p = n$, it can be shown that $0 < \underline{\Psi}^{NC}$ and $\bar{\Psi}^{NC} = \frac{c}{n^2}$. Thus, there exists a range of Ψ such that the grand coalition is stable. Furthermore, if the grand coalition is stable, it Pareto-dominates $p^* = 3$.

For $3 < p < n$, it can be shown that $\frac{c}{n^2} < \underline{\Psi}^{NC} < \bar{\Psi}^{NC}$.[20] (See Online Appendix 3.) That is, the sufficient condition for the existence of a unique equilibrium, $\Psi < \frac{c}{n^2}$, is violated for $sign\left[\Omega^{NC}(p,n)\right] \geq 0$. Thus, there exists no other stable coalition than $p^* = 3$ and $p^* = n$ in the NC-scenario for $n \geq 7$.

We derive the sign of $\Omega^{ST}(p) := W_S^{ST*}(p) - W_{NS}^{ST*}(p-1)$ for the ST-scenario:

$$sign\left[\Omega^{ST}(p,n)\right] = -\left[\Phi_1(p,n)\cdot\Psi^4 - c\cdot\Phi_2(p,n)\cdot\Psi^3 + c^2\cdot\Phi_3\cdot\Psi^2 + c^3\cdot\Phi_4\cdot\Psi + c^4\cdot(p^2 - 4p + 3)\right]$$

with:

$\Phi_1(p,n) = -(n-p+1)^2\cdot(n-p)^2$,

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

$\Phi_3(p,n) = p^4 + (-2n-4)\cdot p^3 + (n^2+8n+9)\cdot p^2 + (-4n^2 - 12n - 12)\cdot p + 2n^2 + 8n + 5$,

[20] The reader will appreciate, by consulting the Online Appendices 3 and 4, that a graphical analysis can be justified in some exceptional cases due to the complexity of some terms. See Eichner and Pethig (2015, 2017) for a similar procedure.

$\Phi_4(p,n) = p^3 \cdot (6-p) + p \cdot (-2np + 8n - 15p + 18) - 6n - 8$ and

$\Psi = \dfrac{f^2 - gd}{d}$. If $sign\left[\Omega^{ST}(p,n)\right] \geq 0$, internal stability holds. We assume, again, $n \geq 7$. Again, we focus on $p \geq 3$ because smaller coalitions are internally stable but not externally stable.

It can be shown that $\Phi_1(p,n) < 0$ for any $p < n$ and $\Phi_1(p,n) = 0$ for $p = n$, $\Phi_2(p,n) > 0$ for any p, $\Phi_3(p,n) \leq 0$ for any p if $n \leq \dfrac{p^3 - 4p^2 + 6p - 4 + \sqrt{p^4 - 3p^2 - 4p + 6}}{p^2 - 4p + 2}$ and $\Phi_3(p,n) > 0$ otherwise and $\Phi_4(p,n) < 0$ for any p.

For $p = 3$, $sign\left[\Omega^{ST}(3,n)\right]$ is a polynomial of degree four, with one zero point at $\Psi = 0$. We solve for the remaining zero points by reducing the function to a polynomial of degree three by dividing by Ψ and using the Cardano formula. (See Online Appendix 4.) For $n \geq 7$, there exists only one remaining zero point, which is negative. It can be shown that $sign\left[\Omega^{ST}(3,n)\right] \geq 0$ for $\Psi < \dfrac{c}{n^2}$, proving that any coalition of size $p = 3$ is internally stable.

For $p = n$, $sign\left[\Omega^{ST}(n,n)\right]$ is a polynomial of degree three, as $\Phi_1(n,n) = 0$. We solve for the zero points of $sign\left[\Omega^{ST}(n,n)\right]$ by using the trigonometric approach for solving cubic equations and obtain three zero points, Ψ_0, Ψ_1 and Ψ_2. (See Online Appendix 4.) It can be shown that $\Psi_2 < 0 < \Psi_0 < \dfrac{c}{n^2} < \Psi_1$ and $sign\left[\Omega^{ST}(n,n)\right] \geq 0$ for $\Psi \in [\underline{\Psi}^{ST}; \overline{\Psi}^{ST}]$ with $\underline{\Psi}^{ST} = \Psi_0$ and $\overline{\Psi}^{ST} = \dfrac{c}{n^2}$, given that we investigate $\Psi > 0$. Thus, there exists a range of Ψ such that the grand coalition is stable. Moreover, $sign\left[\Omega^{ST}(n,n)\right] < 0$ for $\underline{\Psi}^{NC}$, confirming that $\underline{\Psi}^{NC} < \underline{\Psi}^{ST}$. Hence, the range of Ψ for which the grand coalition is stable is smaller in the ST- than NC-scenario. Furthermore, if the grand coalition is stable, it also Pareto-dominates $p^* = 3$ in the ST-scenario.

For $3 < p < n$, $sign\left[\Omega^{ST}(p,n)\right]$ is a polynomial of degree four. We solve for the potential zero points by building the cubic resolvent of the polynomial. This yields eight potential solutions for $sign\left[\Omega^{ST}(p,n)\right]$. It can be graphically shown that these eight solutions Ψ_i, $i \in [1;8]$, are either not an element of $\Psi \in [0;\frac{c}{n^2}]$ or not a zero point of $sign\left[\Omega^{ST}(p,n)\right]$. (See Online Appendix 4.) Thus, $sign\left[\Omega^{ST}(p,n)\right]$ does not change for $\Psi < \frac{c}{n^2}$. It can be shown that $sign\left[\Omega^{ST}(p,n)\right] < 0$ for $\Psi \leq \frac{c}{n^2}$. Thus, there exists no stable coalition other than $p^* = 3$ and $p^* = n$ in the ST-scenario for $n \geq 7$.

A.6 Simulation Strategy

We run a comprehensive number of simulations based on payoff function (10). For each simulation run, we consider payoff functions (10.a) and (10.b) for the same parameter values. (Setting f in (10.a) to $-f$ gives (10.b)). Thus, we cover the case of mitigation and adaptation being substitutes and complements. All parameter values have to satisfy conditions C3 to C5 as explained in Appendix A.4. We start from an initial parameter configuration, and then systematically vary each parameter in order to cover all possible cases.

Downward sloping reaction functions $(\Psi < 0)$

Table 2 in the text considers 17 simulations with different values for the mitigation cost parameter c in order to cover the full range of possible slopes of reaction functions in mitigation space where we display the slope of a single non-signatory's reaction in all tables. That is, we assume the same value for all parameters except for parameter c, and vary c to cover the range $r_{NS}^{'} \in (-1,0)$.

In the Online Appendix 5, three additional tables report results from extensive sensitivity analyses. In Table O.1, we vary the benefit parameter g to cover $r_{NS}^{'} \in (-1,0)$ keeping other parameters fixed. We consider 15 different values of parameter g.

Table O.2 and Table O.3 in Online Appendix 5 selects those parameter constellations from Table 2 in the text and Table O.1 in the Online Appendix for which $p^{ST*} \geq 80$ (given $n = 100$) as starting values and varies other parameters. In Table O.2, we vary the benefit parameters b and β by considering 48 different parameter combinations, and in Table O.3 we vary parameters f and d considering 38 different parameter combinations where d is the adaptation cost parameter and f measures the marginal decrease (increase) of the marginal benefit from adaptation due to an increase in total mitigation. Changes of parameters b and β do not affect the slope of the reaction function in mitigations space and, hence, also not p^{ST*}, whereas changes of parameters f and d affect the slope and, hence also p^{ST*}. Table O.2 and O.3 serve to test the robustness of the conclusion regarding the paradox of cooperation by considering large stable coalitions.

Upward sloping reaction functions $(\Psi > 0)$

In case of upward sloping reaction functions in mitigation, we face a constraint on the upper bound of the slope of reaction functions. See Appendix A.1. Only flat reaction functions do not violate the condition for the existence of a unique interior equilibrium.

In Table 3 in the text, various combinations of parameters g, f and d (that determine the value of Ψ) and of parameter c are considered in order to cover all possible outcomes in terms of stable coalition sizes, as stated in Proposition 3. (All four parameters affect the slope of the reaction function in mitigation space.) 15 simulations are reported. Additional sensitivity analyses are conducted in Table O.4 and O.5 in Online Appendix 5.

Table O.4 focuses on those cases of Table 3 in the text for which the grand coalition is stable in both scenarios, the NC- and ST scenario. For each case, we vary parameters g, f and d in order to show that $p^{*}=n$, for any given level of the cost parameter c, can be achieved. According to Proposition 3, $\underline{\Psi} \leq \Psi$ is required to have $p^{*}=n$. Eight cases in which the grand coalition is stable in both scenarios are reported in Table O.4.

Finally, Table O.5 focuses on the simulations for which $p^{*}=n$ in Table 3, either in NC- or in both scenarios, and performs a sensitivity analysis with respect to the benefit parameters b and β in order to test the robustness of our conclusions regarding the paradox of cooperation. Thirty different combinations are reported.

Tables

Table 1: First Order Conditions under the NC- and ST-Scenario*

	NC-scenario	ST-scenario
Signatories	$p \cdot B_M\left(M, a_i\right) = C_m\left(m_S\right)$ (3.a)	$p \cdot \left[B_M\left(M, a_i\right)\left(1 + R_{NS}^{'}\right)\right] = C_m\left(m_S\right)$ (5.a)
Non-Signatories	$B_M\left(M, a_i\right) = C_m\left(m_{NS}\right)$ (3.b)	$B_M\left(M, a_i\right) = C_m\left(m_{NS}\right)$ (5.b)
Both	$B_a\left(M, a_i\right) = D_a\left(a_i\right)$ (4)	$B_a\left(M, a_i\right) = D_a\left(a_i\right)$ (6)

* Let $M_{NS} = R_{NS}(M_S)$. Then, $R_{NS}^{'} = {}^{\partial M_{NS}}\!/_{\partial M_S}$ with $M_S = p \cdot m_S$ and $M_{NS} = (n - p) \cdot m_{NS}$. For further details on $R_{NS}^{'}$ see the discussion below, in particular equation (8).

Simulations for $\Psi < 0$, implying downward sloping reaction functions in mitigation space

Table 2: Different slopes of reaction functions through a variation of parameter c*, #

SIMULATIONS	r'_{NS}	Ψ	NASH-COURNOT						STACKELBERG						Substitutability				Complementarity			
			SAD	PEP	PIP	MCOH	WCOH	p*	SAD	PEP	PIP	MCOH	WCOH	p*	*h'(M)*	ICI	INI NC	INI ST	*h'(M)*	ICI	INI NC	INI ST
c=500	-0.0016	-0.80	✓	✓	✓	✓	✓	2	✓	✓	✓	✓	✓	3	-0.20	135.57	0.79	1.89	0.20	186.92	1.09	2.60
c=300	-0.0027	-0.80	✓	✓	✓	✓	✓	2	✓	✓	✓	✓	✓	3	-0.20	94.19	0.82	1.69	0.20	121.10	1.06	2.17
c=150	-0.0053	-0.80	p>10	✓	✓	✓	✓	1	✓	✓	✓	✓	✓	3	-0.20	49.76	0	1.02	0.20	59.87	0	1.23
c=50	-0.0157	-0.80	p>15	✓	p>6	✓	✓	1	✓	✓	p>2	✓	✓	3	-0.20	13.10	0	0.09	0.20	14.99	0	0.09
c=20	-0.0385	-0.80	p>15	✓	p>8	✓	✓	1	✓	p>2	p>4	p>2	p>2	5	-0.20	3.24	0	0.01	0.20	3.66	0	0.01
c=10	-0.0741	-0.80	p>16	✓	p>9	✓	✓	1	✓	p>4	p>8	p>4	p>4	9	-0.20	0.98	0	0.01	0.20	1.10	0	0.02
c=5	-0.1379	-0.80	p>17	✓	p>9	✓	✓	1	✓	p>7	p>14	p>7	p>7	15	-0.20	0.27	0	0	0.20	0.31	0	0
c=3	-0.2105	-0.80	p>17	✓	p>9	✓	p>12	1	✓	p>11	p>21	p>11	p>11	22	-0.20	0.10	0	0	0.20	0.12	0	0
c=1.5	-0.3478	-0.80	p>17	✓	p>9	✓	p>14	1	✓	p>19	p>35	p>19	p>18	36	-0.20	0.03	0	0	0.20	0.03	0	0
c=1	-0.4444	-0.80	p>17	✓	p>9	✓	p>15	1	✓	p>25	p>44	p>25	p>24	46	-0.20	0.01	0	0	0.20	0.01	0	0
c=0.7	-0.5333	-0.80	p>17	✓	p>9	✓	p>15	1	✓	p>31	p>53	p>31	p>29	54	-0.20	0.01	0	0	0.20	0.01	0	0
c=0.5	-0.6154	-0.80	p>17	✓	p>9	✓	p>15	1	✓	p>38	p>61	p>38	p>34	62	-0.20	0	0	0	0.20	0	0	0
c=0.3	-0.7273	-0.80	p>17	✓	p>9	✓	p>15	1	✓	p>47	p>72	p>47	p>42	74	-0.20	0	0	0	0.20	0	0	0
c=0.2	-0.8000	-0.80	p>17	✓	p>9	✓	p>16	1	✓	p>55	p>79	p>55	p>49	81	-0.20	0	0	0	0.20	0	0	0
c=0.1	-0.8889	-0.80	p>17	✓	p>9	✓	p>16	1	✓	p>66	p>89	p>66	p>59	90	-0.20	0	0	0	0.20	0	0	0
c=0.05	-0.9412	-0.80	p>17	✓	p>9	✓	p>16	1	✓	p>75	p>94	p>75	p>67	95	-0.20	0	0	0	0.20	0	0	0
c=0.005	-0.9938	-0.80	p>17	✓	p>9	✓	p>16	1	✓	p>92	p>99	p>92	p>87	100	-0.20	0	0	0	0.20	0	0	0

* For the other parameters we assume throughout: b=10, β=10, g=1, f=1, d=5.

Remark: Lowering parameter c increases the absolute value of the slope of the reaction function in mitigation space.

# ✓ means this property holds for all values of p, $p > x$ means this property holds for all values of p larger than x. Headings and abbreviations as explained in the text.

If SAD holds for a given *p*, it means that the move from *p-1* to *p* is superadditive.

The values of Ψ, ICI and INI are rounded to 2 digits and the values of r'_{NS} are rounded to 4 digits.

Simulations for $\Psi > 0$, implying upward sloping reaction functions in mitigation space

Table 3: Variation of parameters f, g and d that affect the value of Ψ and the cost parameter c to cover different sizes of stable coalitions[*, #]

SIMULATIONS	r'_{NS}	Ψ	NASH-COURNOT						STACKELBERG						Substitutability				Complementarity			
			SAD	PEP	PIP	MCOH	WCOH	p*	SAD	PEP	PIP	MCOH	WCOH	p*	$h'(M)$	ICI	INI NC	INI ST	$h'(M)$	ICI	INI NC	INI ST
Base simulation	0	0.11	✓	✓	✓	✓	✓	3	✓	✓	✓	✓	✓	3	-0.33	176.42	0.21	0.21	0.33	618.04	0.73	0.73
g=1.51	0	0.60	✓	✓	✓	✓	✓	3	✓	✓	✓	✓	✓	3	-0.33	196.45	0.21	0.21	0.33	688.13	0.73	0.73
g=2.11	0	0	✓	✓	✓	✓	✓	3	✓	✓	✓	✓	✓	3	-0.33	172.47	0.21	0.21	0.33	604.19	0.72	0.72
f=6.33	0	0	✓	✓	✓	✓	✓	3	✓	✓	✓	✓	✓	3	-0.32	176.71	0.21	0.21	0.32	597.52	0.72	0.72
f=6.99	0	0.44	✓	✓	✓	✓	✓	3	✓	✓	✓	✓	✓	3	-0.35	176.39	0.19	0.19	0.35	685.29	0.75	0.75
c=45001	0	0.11	✓	✓	✓	✓	✓	3	✓	✓	✓	✓	✓	3	-0.33	195.76	0.23	0.23	0.33	679.23	0.79	0.80
c=100000	0	0.11	✓	✓	✓	✓	✓	3	✓	✓	✓	✓	✓	3	-0.33	88.74	0.11	0.11	0.33	325.39	0.39	0.39
d=6.51	0.0001	4.49	✓	✓	✓	✓	✓	3	✓	✓	✓	✓	✓	3	-0.99	0	0	0	0.99	4604.72	0.57	0.58
d=21.1	0	0	✓	✓	✓	✓	✓	3	✓	✓	✓	✓	✓	3	-0.31	190.46	0.23	0.23	0.31	619.07	0.74	0.74
CASE 1**	0.0001	4.99	✓	✓	✓	✓	✓	100	✓	✓	✓	✓	✓	3	-0.99	0.01	0.01	0	0.99	1.23x10^{6}	1.23x10^{6}	0.62
CASE 2**	0.0001	4.99	✓	✓	✓	✓	✓	100	✓	✓	✓	✓	✓	100	-0.99	0	0	0	0.99	1.26x10^{9}	1.26x10^{9}	1.26x10^{9}
CASE 3**	0.0001	0.09	✓	✓	✓	✓	✓	100	✓	✓	✓	✓	✓	100	-0.19	0	0	0	0.19	3.02x10^{6}	3.02x10^{6}	3.02x10^{6}
CASE 4**	0.0001	0.99	✓	✓	✓	✓	✓	100	✓	✓	✓	✓	✓	100	-0.99	0	0	0	0.99	1.58x10^{7}	1.58x10^{7}	1.58x10^{7}
CASE 5**	0.0001	1.99	✓	✓	✓	✓	✓	100	✓	✓	✓	✓	✓	100	-0.99	0	0	0	0.99	2.46x10^{6}	2.46x10^{6}	2.46x10^{6}
CASE 6**	0.0001	9.99	✓	✓	✓	✓	✓	100	✓	✓	✓	✓	✓	3	-0.99	0.01	0.01	0	0.99	1.31x10^{6}	1.31x10^{6}	0.66

* For the base simulation, we assume: b=10, β=10, g=2, f=6.5, c=50000, d=5.

** For the six different cases for which $p^*=n$ in the NC and ST scenario, we assume:
CASE 1: b=10, β=10, g=2, f=6.9989, c=50000, d=7;
CASE 2: b=10, β=10, g=2, f=6.9999, c=50000, d=7;
CASE 3: b=1, β=5, g=0.3, f=1.9999, c=1000, d=10;
CASE 4: b=10, β=10, g=21, f=21.9999, c=10000, d=22;
CASE 5: b=10, β=10, g=0.0001, f=1.9999, c=20000, d=2;
CASE 6: b=10, β=10, g=5, f=14.9979, c=100000, d=15.

Remark: Note that CASE 1 and CASE 2 are obtained from the base simulation, changing both parameters f and d.

[#] See Table 2.