CESifo WORKING PAPERS

11040
2024

March 2024

How Important Are IEAs for Mitigation If Countries Are of the Homo Moralis Type?

Thomas Eichner, Rüdiger Pethig

CESifo

Impressum:

CESifo Working Papers
ISSN 2364-1428 (electronic version)
Publisher and distributor: Munich Society for the Promotion of Economic Research - CESifo GmbH
The international platform of Ludwigs-Maximilians University's Center for Economic Studies and the ifo Institute
Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740, Telefax +49 (0)89 2180-17845, email office@cesifo.de
Editor: Clemens Fuest
https://www.cesifo.org/en/wp

An electronic version of the paper may be downloaded
· from the SSRN website: www.SSRN.com
· from the RePEc website: www.RePEc.org
· from the CESifo website: https://www.cesifo.org/en/wp

CESifo *Working Paper No. 11040*

How Important Are IEAs for Mitigation If Countries Are of the Homo Moralis Type?

Abstract

We analyze international environmental agreements in a two-stage game when governments have homo moralis preferences à la Alger and Weibull (2013, 2016). The countries base their decisions on the material payoff obtained on the hypothesis that all other countries act as they with predetermined probability. They are assumed to act morally w.r.t. both membership and emissions. We investigate the interaction and impact of that moral behavior on coalition formation and material payoff. The membership morality tends to increase while the emissions morality tends to decrease the coalition size, but the outcome is not smoothly determined by these opposite forces.

JEL-Codes: C720, Q500, Q580.

Keywords: IEA, stability, homo moralis, emissions morality, membership morality.

Thomas Eichner
Department of Economics
University of Hagen / Germany
thomas.eichner@fernuni-hagen.de

Rüdiger Pethig
Department of Economics
University of Siegen / Germany
pethig@vwl.wiwi.uni-siegen.de

Financial support from the German Science Foundation (DFG grant number EI 847/2-1) is gratefully acknowledged.

1 Introduction

It is a well-known result in theoretical environmental economics that all countries incur large material welfare losses, if they cope with man-made problems of global environmental deterioration such as climate change in a non-cooperative and purely self-interested fashion. Some studies in the early 1990s (to be reviewed below) analyze in a parsimonious game how that pessimistic result may be improved upon through cooperation. In that benchmark game of the literature countries first decide on their membership in an International Environmental Agreement (IEA) and then on the emissions of a global pollutant. Due to the assumption that countries are of the *homo-oeconomicus* type, the free-rider incentives turn out to be so strong that the IEA consists of no more than three signatories.

In the present paper we depart from that basic IEA game (but reproduce it as a special case) by assuming that the countries are guided by moral considerations, to some extent at least. We conceive of the countries' moral behavior as being induced by their constituencies' moral stance on domestic emissions and international cooperation. In recent years, increasing numbers of individuals deliberately reduced their carbon footprint below the level of self-interested consumers (Ellen et al. 2013 and Liobikiene et al. 2016). It appears to be realistic, therefore, that in their role as voters they induce governments to play a more pro-active part in emissions reduction and global cooperation.

Here we focus on morality in the spirit of Kantian ethics (Kant 1785) the core of which is the categorical imperative. It says that one should take those actions and only those actions that one would advocate all others take as well. Kantian behavior focuses on "doing the right thing" and thus differs from both self-interested and altruistic behavior. We will apply the specific formalization of moderate moral behavior suggested by Alger and Weibull (2013, 2016, 2017, 2020) which they refer to as *homo moralis* behavior. Their *homo moralis* is an individual who maximizes her material payoff on the hypothesis that all other individuals choose the same action(s) as her with predetermined probability. That probability is referred to as degree of morality $\kappa \in [0, 1]$, where the polar cases $\kappa = 0$ and $\kappa = 1$ specify *homo oeconomicus* behavior and *homo kantiensis* behavior, respectively. The concept of *homo moralis* behavior with degree of morality $\kappa \in]0, 1[$ has strong theoretical appeal (to us), because of Alger and Weibull's (2013, 2016) deep result that populations with that behavior are evolutionary stable. From a theoretical point of view our paper is the first that applies the morality concept of Alger and Weibull (2013, 2016) to two-stage games with moral behavior on both stages.

The present paper aims to investigate the solution of the two-stage IEA game with countries of the homo moralis type with respect to both the size of the IEA (called coalition, henceforth) and the countries' material welfares. Accordingly, we need to introduce two different kinds of degrees of morality: κ_ε for the morality with respect to emissions (ε-morality) and κ_μ for the morality with respect to membership (μ-morality). Technically speaking, we will characterize (i) the mapping from the set of feasible morality parameters $(\kappa_\varepsilon, \kappa_\mu)$ to the equilibrium coalition size and (ii) the mapping from the same domain to the countries' equilibrium material payoffs. Our IEA game yields the following new results:

- In the absence of μ-morality, the three-country coalition of the scenario $\kappa_\varepsilon = \kappa_\mu = 0$ is destabilized such that from some small positive κ_ε onward there is no coalition anymore.
- In the absence of ε-morality, increasing κ_μ increases the coalition size progressively such that the grand coalition and with it the first-best material payoff, are attained at all κ_μ greater than some intermediate level of κ_μ.
- Unless the parameters $(\kappa_\varepsilon, \kappa_\mu)$ yield either no coalition or the grand coalition, it is not possible to keep the coalition size constant by reducing κ_ε and increasing κ_μ (or vice versa) in small amounts; in that sense, μ-morality and ε-morality are no substitutes, neither with respect to the coalition size nor with respect to the material payoff.
- For most $(\kappa_\varepsilon, \kappa_\mu)$ satisfying $\kappa_\mu/\kappa_\varepsilon \leq 1$ the ε-morality dominates the μ-morality such that there exists no coalition; the material payoff is increasing in κ_ε and reaches the first-best level at $\kappa_\varepsilon = 1$.
- For a large set of $(\kappa_\varepsilon, \kappa_\mu)$ satisfying $\kappa_\mu/\kappa_\varepsilon > 1$ the μ-morality dominates the ε-morality such that the grand coalition forms with the material payoff being at its first-best level.
- If $\kappa_\mu = \kappa_\varepsilon = \kappa$ is at a low to intermediate level, there exists no stable coalition, but the material payoff increases in κ; at some intermediate level of κ the coalition size steeply jumps up to its maximum and stays there for all higher levels of κ; so, for low [high] levels of κ the emissions [membership] morality fully determines the outcome.

The present paper is related to Ulph and Ulph (2023) and Eichner and Pethig (2024) who also study two-stage IEA games with moral governments. They construct the countries' moral preferences as a convex combination of *homo oeconomicus* preferences and *homo kantiensis* preferences. Moral decisions are taken only in the emissions game in stage 2, whereas in the two-stage game of the present paper the countries exhibit μ-moral behavior at state 1 and ε-moral behavior at stage 2. The approaches of Ulph and Ulph (2023) and Eichner and Pethig (2024) differ from each other with respect to the specification of moral preferences, and both of them differ from the *homo moralis* approach of the present paper. A major difference between our present approach and their approaches is that they find μ-morality and ε-morality to be substitutes while we do not.

The present paper contributes to the literature on IEAs and to the literature on morality and Kantian behavior. The seminal literature on IEAs when countries play Nash[1] goes back to Hoel (1992), Carraro and Siniscalco (1991, 1993) and Finus and Maus (2008). These studies apply parametric functional forms with quadratic benefits and linear or quadratic environmental damages. If damages are quadratic, the coalition consists of two countries. If damages are linear, the coalition consists of three countries.[2] Governments are assumed to be purely self-interested, they display the human image of *homo oeconomicus* which predominates in economics. Deviations from *homo oeconomicus* behavior in the context of coalition formation can be found in Lange and Vogt (2003), Van der Pol et al. (2012), Vogt (2016), Nyborg (2018), Buchholz et al. (2018) and Schopf (2023). Van der Pol et al. (2012) and Schopf (2023) study the impact of altruism on coalition formation. Van der Pol et al. (2012) find that altruism with respect to membership results in large coalitions, whereas Schopf (2023) points out that coalitions are small when countries are altruistic with respect to both emissions and membership. Nyborg (2018) and Buchholz et al. (2018) elaborate that reciprocity preferences may increase the participation in a coalition. Lange and Vogt (2003) and Vogt (2016) consider the formation of coalitions when countries are inequality averse. Lange and Vogt (2003) find that inequality aversion as proposed by Bolton and Ockenfels (2000) can increase the coalition size, whereas Vogt (2016) shows that inequality aversion as proposed by Fehr and Schmidt (1999) disappoints the expectations of large coalitions.

As to the literature on moral preferences, Laffont (1975) was the first who introduced Kantian behavior in a formal way. He considers a Kantian agent who maximizes her utility on the counterfactual assumption that other agents act as her. The above mentioned convex combination of *homo oeconomicus* preferences and *homo kantiensis* preferences has been applied by Daube and Ulph (2016) and Eichner and Pethig (2021) in the context of environmental regulation and international (non-cooperative) climate policy, respectively.

Alger and Weibull (2017, 2020) applied their own homo moralis approach to public good provision, to environmental economics and to coordination issues. Alger and Laslier (2022) applied that approach to voting. However, to the best of our knowledge, their ap-

[1] There is another strand of the IEA literature that considers Stackelberg games in which the coalition acts as Stackelberg leader and the fringe countries act as Stackelberg followers (Barrett 1994, Rubio and Ulph 2006 and Diamantoudi and Sartzetakis 2006). In case of linear damages, the outcome of Nash and Stackelberg games is the same.

[2] We reproduce that result in the present paper for the special case $\kappa_\varepsilon = \kappa_\mu = 0$.

proach has not yet been employed in two-stage IEA games with the special feature that governments act morally with respect to both membership and emissions. A different notion of Kantian behavior was developed by Roemer (2010, 2015) with an equilibrium concept called Kantian equilibrium. Roemer's Kantian behavior has been taken up by Grafton et al. (2017) and Van Long (2020). Compared to *homo oeconomicus* behavior that kind of Kantian behavior is shown to reduce inefficiencies in case of public goods and externalities.

Finally, there is a small literature analyzing moral or pro-environmental behavior that is not explicitly linked to Kant. Herweg and Schmidt (2022) investigate how emissions taxes and cap-and-trade schemes affect moral behavior. In Ambec and De Donder's (2022) public choice approach, consumers who differ with respect to their warm glow from purchasing environmentally friendly goods vote on taxes and standards. Aghion et al. (2023) study how environmental concerns of consumers affect innovation in clean technologies.

The paper is organized as follows. Section 2 describes the building blocks of the basic IEA game without moral countries and explicitly formalizes the membership decisions for use in the subsequent analysis. Section 3 analyzes the emissions game at stage 2 with ε-moral countries when the membership decisions are given. In section 4 we study the membership game at stage 1 of μ-moral countries who anticipate their ε-moral decisions on emissions at stage 2. Section 5 determines the properties of the solution of the two-stage IEA game. It analyzes the scenario with ε-morality only in section 5.1, the scenario with μ-morality only in section 5.2, and the full scenario with both ε - and μ-morality in section 5.3. Section 6 concludes.

2 Players, strategies and payoffs

The players in the two-stage IEA game are $n \geq 3$ ex ante identical countries. Each of them emits a pollutant that generates benefits for the emitting country (only) and contributes to the environmental damage caused by the aggregate emissions of all countries. The countries' options are to determine their emissions cooperatively in an environmental coalition or non-cooperatively. In stage 1 of the game, each country decides whether it wants to become a member of such a coalition. In stage 2 the countries within and outside the coalition choose the level of emissions.

In formal terms, country i's membership decision is the action $s_i \in \{0, 1\}$, where $s_i = 1$ is the decision to be a member of the coalition and $s_i = 0$ is the decision to stay outside the coalition. Every membership profile $\mathbf{s} = (s_1, s_2, \ldots, s_n) \in \Psi = \big\{\mathbf{s} \big| s_i \in \{0, 1\}, i = 1, \ldots, n\big\}$

is characterized by the sets $C(\mathbf{s})$ and $F(\mathbf{s})$ of coalition countries and fringe countries, respectively, and by the coalition size[3] $m := \sum_{j=1}^{n} s_j$. Due to our assumption that all countries are ex ante identical, we can disregard the order in the profile $(s_1, s_2, \ldots, s_n)$. The only information each country needs to know about a profile $\mathbf{s}$ is its own decision $s_i \in \{0, 1\}$ and the implied coalition size $m = \sum_{j=1}^{n} s_j$. By $e_i \in \mathbb{R}_+$ we denote country i's emissions, $i = 1, \ldots, n$, and by $\mathbf{e} = (e_1, e_2, \ldots, e_n) \in \mathbb{R}_+^n$ we denote the vector of emissions. Since the membership decision country i made in stage 1 matters for its decision on emissions in stage 2 (in a way to be specified later), it is convenient to denote the action e_i by e_i^f, if $s_i = 0$, and by e_i^c, if $s_i = 1$. We assume that if a membership profile $\mathbf{s} \in \Psi$ with coalition size $\sum_{j=1}^{n} s_j = m$ is given, country i's material payoff is equal to[4]

$$\Pi(\mathbf{e}, m) = B(e_i) - D\left(\sum_{j \in C(\mathbf{s})} e_j + \sum_{j \in F(\mathbf{s})} e_j\right) = \begin{cases} \Pi^c\left(e_i^c, e_a^c, e_a^f, m\right) & \forall\, i \in C(\mathbf{s}), \\ \Pi^f\left(e_i^f, e_a^c, e_a^f, m\right) & \forall\, i \in F(\mathbf{s}), \end{cases} \tag{1}$$

where

$$\Pi^c\left(e_i^c, e_a^c, e_a^f, m\right) = B(e_i^c) - D\left[e_i^c + (m-1)e_a^c + (n-m)e_a^f\right], \tag{2a}$$

$$\Pi^f\left(e_i^f, e_a^c, e_a^f, m\right) = B(e_i^f) - D\left[e_i^f + (n-m-1)e_a^f + m e_a^c\right]. \tag{2b}$$

$B(\cdot)$ are country i's benefits from own emissions e_i^h, $h = c, f$, and $D(\cdot)$ is the environmental damage caused by the sum of all countries' emissions. Function B satisfies $B' > 0$, $B'' < 0$ and function D satisfies $D' > 0, D'' \geq 0$. The payoffs (1), (2a) and (2b) are written such that the focus is on the emissions decision of a country i that is either a member of the coalition or fringe. We take advantage of symmetry by assuming that apart from country i under consideration all countries in the same group choose the same level of emissions, but these emissions may and will differ across groups. Technically, that assumption is formalized by the subscripts a in (1), (2a) and (2b). If $i \in C(\mathbf{s})$, we set $e_a^c = e_j^c$ for all $j \in F(\mathbf{s})$, $j \neq i$, and $e_a^c = e_j^c$ for all $j \in C(\mathbf{s})$, $j \neq i$.

Throughout the paper we will specify the functions B and D from (1) by the parametric functional forms

$$B(e_i) = \alpha e_i - \frac{\beta}{2} e_i^2 \quad \text{and} \quad D\left(\sum_{j=1}^{n} e_j\right) = \delta \sum_{j=1}^{n} e_j, \tag{3}$$

[3] Throughout the paper we write m for $\sum_{j=1}^{n} s_j$ to avoid clumsy notation.

[4] All functions defined in (1) and (2a) and (2b) also depend on the number of countries, n. We suppress n as an argument here and in other functions below, however, unless n is analytically relevant, as in the Appendix.

where α, β and δ are positive parameters. For the benefit of obtaining informative results, that kind of parametrization is widely applied in the literature on IEA games without moral countries. The linearity of the damage function provides particularly strong analytical relief, because it results in dominant strategies.

So far, the description of the IEA game is that of the standard game in the environmental economics literature on IEA games without moral countries. Now we turn to the crucial difference between the IEA games with and without morality which is how the decisions on emissions and membership are made. We follow the usual procedure to first describe and solve the emissions game in stage 2 of the IEA game (Section 3) and after that we analyze the stage 1 membership game (Section 4).

3 The stage 2 emissions game with ε-moral countries

In the standard IEA game of the literature, the usual assumption is that fringe countries maximize the material payoff (2b) taking the emissions of all other countries as given (Nash behavior). The coalition countries produce an environmental agreement by committing to a collective decision on their emissions. The emissions they choose maximize the sum of all coalition countries' material payoffs (2a) taking the emissions of all fringe countries as given. All countries are purely self-interested in the sense that they only care about their own material payoff. This is obviously true for fringe countries, but also for coalition countries insofar as they join the coalition only if their material payoff is (weakly) higher in the coalition than in the fringe.

If countries are moral their material payoff (1) remains a relevant concept, but their decisions are based on a decision function that differs from their material payoff. Specifically, we apply Alger and Weibull's (2013, 2016, 2017, 2020) concept of homo moralis preferences to the countries in our model having in mind that the governments' moral actions are induced by (sufficiently many) individuals with homo moralis preferences in their constituency. Alger and Weibull's homo moralis is an individual who acts on the hypothesis that all other individuals choose the same action as her with some predetermined probability $\kappa \in [0, 1]$. So, the decision function of homo moralis is not her material payoff, but rather that material payoff which results, if all other individuals *would* choose the same level of emissions as it with probability κ. A well-known version of Immanuel Kant's (1785, p. 30) categorical imperative is that you should "act only according to that maxim whereby you can at the same time will that it should become a universal law." Alger und Laslier (2022, p. 283) aptly observe that "... Homo moralis can be said to "act according to that maxim whereby you can at the same time will that others should do likewise with some probability"" (here: probability $\kappa \in [0,1]$). This interpretation describes Kant's categorical imperative as the extreme case $\kappa = 1$ of homo moralis, and the opposite polar case $\kappa = 0$ describes the purely self-interested homo oeconomicus who is prominent in conventional economics.

In analogy to homo moralis, we consider countries to be of the homo moralis type, if they determine their emissions on the hypothesis that all other coalition and fringe countries choose the same level of emissions as them with some predetermined probability. We denote that probability as $\kappa_\varepsilon \in [0,1]$ and call it the degree of ε-morality. The subscript ε in κ_ε indicates that the morality under review relates to moral behavior with respect to emissions. A subscript such as ε is necessary, because in stage 1 the countries will be assumed to exhibit moral behavior with respect to membership in addition to their ε-morality.

A country i with degree of morality κ_ε cares about the material payoff it would obtain if all other countries would choose the same emissions as it with probability κ_ε. In the following, we will construct that hypothetical payoff in several steps, and we begin with illustrating country i's calculus with an example.

Example 1. Suppose there are 4 countries, and the membership decisions[5] $(s_i = 1, s_1 = 1, s_2 = 0, s_3 = 0)$ taken at stage 1 are known at stage 2 so that the corresponding 'true' emissions profile is $\left(e_i^c, e_1^c, e_2^f, e_3^f\right)$. If country i hypothesizes that the other countries choose the same emissions as it with probability κ_ε, it faces a set of emissions profiles with certain probabilities in which the emissions of $q = 0, 1, 2, 3$ of the $n - 1 = 3$ other countries are replaced by its own emissions e_i^c. Specifically, it faces

- $\left(e_i^c, e_1^c, e_2^f, e_3^f\right)$ with probability $(1-\kappa_\varepsilon)^3$, if $q = 0$,
- $\left(e_i^c, e_i^c, e_2^f, e_3^f\right)$, $\left(e_i^c, e_1^c, e_i^c, e_3^f\right)$, $\left(e_i^c, e_1^c, e_2^f, e_i^c\right)$ each with probability $\kappa_\varepsilon(1-\kappa_\varepsilon)^2$, if $q = 1$,
- $\left(e_i^c, e_i^c, e_i^c, e_3^f\right)$, $\left(e_i^c, e_i^c, e_2^f, e_i^c\right)$, $(e_i^c, e_1^c, e_i^c, e_i^c)$ each with probability $\kappa_\varepsilon^2(1-\kappa_\varepsilon)$, if $q = 2$,
- $(e_i^c, e_i^c, e_i^c, e_i^c)$ with probability κ_ε^3, if $q = 3$.

We readily generalize the example by observing that if there are $n \geq 3$ countries, an individual emissions profile in which country i replaced the emissions of q out of the other $n-1$ countries appears with probability $\kappa_\varepsilon^q (1-\kappa_\varepsilon)^{n-1-q}$. Since any number of replacements $q \in \{0, 1, \ldots, n-1\}$ can be attained by $\binom{n-1}{q}$ different emissions profiles, the probability of

[5] We could take any other membership profile in our example, but we need to specify such a profile to assign a coalition or fringe membership to each country.

an emissions profile with q replacements is

$$\kappa_{\varepsilon}^{q}\left(1-\kappa_{\varepsilon}\right)^{n-1-q}\binom{n-1}{q}. \tag{4}$$

Summing the probabilities (4) over q yields $\sum_{q=0}^{n-1}\kappa_{\varepsilon}^{q}\left(1-\kappa_{\varepsilon}\right)^{n-1-q}\binom{n-1}{q}=1$.

Country i is not indifferent with respect to alternative emissions profiles which exhibit one and the same replacement number q, because it rightly anticipates that the aggregate emissions implied by these profiles differ. To make those differences precise, suppose $q \in \{0,1\ldots,n-1\}$ is fixed and fringe [coalition] country i replaces the emissions of r other fringe countries [coalition countries] and the emissions of $(q-r)$ coalition countries [fringe countries] by its own emissions. That pattern of replacements yields the material payoff

$$\tilde{\Pi}^{h2}\left(e_i^h, e_a^c, e_a^f, r, q, m\right) := B\left(e_i^h\right) - D\left[E^h(e_i^h, e_a^c, e_a^f, r, q, m)\right] \tag{5}$$

for $h = c, f$, where

$$\begin{aligned} E^c\left(e_i^c, e_a^c, e_a^f, r, q, m\right) &:= (1+q)e_i^c + (m-1-r)e_a^c + (n-m-q+r)e_a^f \quad \forall\, i \in C(\mathbf{s}), \text{(6a)} \\ E^f\left(e_i^f, e_a^c, e_a^f, r, q, m\right) &:= (1+q)e_i^f + (n-m-1-r)e_a^f + (m-q+r)e_a^c \quad \forall\, i \in F(\mathbf{s}). \text{(6b)} \end{aligned}$$

Next suppose country i replaces the emissions of some given $q \in \{0,1,\ldots,n-1\}$ other countries by its own emissions in all feasible ways. Its associated expected material payoff is then equal to

$$\begin{aligned} \hat{\Pi}^{h2}\left(e_i^h, e_a^c, e_a^f, q, m\right) &:= \sum_{r \in R^h} A^h(r,q,m) \cdot \tilde{\Pi}^{h2}\left(e_i^h, e_a^c, e_a^f, r, q, m\right) \\ &= B\left(e_i^h\right) - \sum_{r \in R^h} A^h(r,q,m) \cdot D\left[E^h\left(e_i^h, e_a^c, e_a^f, r, q, m\right)\right] \qquad (7) \end{aligned}$$

for $h = c, f$, where

$$A^c(r,q,m) := \frac{\binom{m-1}{r}\cdot\binom{n-m}{q-r}}{\binom{n-1}{q}}, \quad A^f(r,q,m) := \frac{\binom{n-m-1}{r}\cdot\binom{m}{q-r}}{\binom{n-1}{q}}, \tag{8}$$

$$R^c := \left\{r \in \mathbb{N} \middle| m-1 \geq r \geq q-n+m\right\}, \quad R^f := \left\{r \in \mathbb{N} \middle| n-m-1 \geq r \geq q-m\right\}. \tag{9}$$

$A^h(r,q,m)$, $h = c, f$, is the probability that a coalition country $i \in C(\mathbf{s})$ or fringe country $i \in F(\mathbf{s})$ replaces by its own emissions the emissions of q countries from among the $n-1$ other countries on condition that $r \in R^h$ of these q countries are from its own group and $q-r$ from the other group. $\binom{m-1}{r}$ gives the number of ways that coalition country i can replace by its own emissions the emissions of r coalition countries from among the $m-1$ other coalition countries and $\binom{n-m}{q-r}$ gives the number of ways that coalition country i can replace by its own emissions the emissions of $q-r$ fringe countries from among the $n-m$ fringe countries. The interpretation of the terms $\binom{n-m-1}{r}$ and $\binom{m}{q-r}$ is analogous. By definition of $A^h(r,q,m)$ it holds that $\sum_{r\in R^h} A^h(r,q,m) = 1$.

The probability $\kappa_\varepsilon^q (1-\kappa_\varepsilon)^{n-1-q} \binom{n-1}{q}$ from (4) which we established to be the probability that an emissions profile occurs with $q \in \{0,1,\ldots,n-1\}$ replacements obviously is also the probability that the payoff is equal to $\hat{\Pi}^{h2}\left(e_i^h, e_a^c, e_a^f, q, m\right)$ from (7). Accounting for all feasible values of q it follows that the expected material payoff a coalition or fringe country expects on the hypothesis that all other countries would choose the same emissions as it with probability κ_ε is equal to

$$\begin{aligned} W^h\left(e_i^h, e_a^c, e_a^f, m, \kappa_\varepsilon\right) &:= \sum_{q=0}^{n-1} \kappa_\varepsilon^q (1-\kappa_\varepsilon)^{n-1-q} \binom{n-1}{q} \cdot \hat{\Pi}^{h2}(e_i^h, e_a^c, e_a^f, q, m) \\ &= B(e_i^h) - \sum_{q=0}^{n-1} \kappa_\varepsilon^q (1-\kappa_\varepsilon)^{n-1-q} \binom{n-1}{q} \cdot \sum_{r\in R^h} A^h(r,q,m) \cdot D[E^h(e_i^h, e_a^c, e_a^f, r, q, m)] \end{aligned} \quad (10)$$

for $h = c, f$. The equation (10) defines the payoff functions constructed according to Alger and Weibull's homo moralis concept and applied to the stage 2 emissions game. We refer to W^h, $h = c, f$, in (10) as country i's *ε-moral welfare* and assume that the countries apply these welfare functions as decision functions in the stage 2 emissions game. In that game, the coalition countries' choice of emissions is a collective decision. The coalition acts as (if it is) a single player who may be referred to as the manager of the coalition. That manager's payoff is the sum of the coalition countries' ε-moral welfares, $\sum_{j\in C(\mathbf{s})} W^c\left(e_j^c, e_a^c, e_a^f, \kappa_\varepsilon, m\right)$, and she chooses all coalition countries' emissions $e_j^c \in \mathbb{R}_+$ for all $j \in C(\mathbf{s})$. The fringe country $i \in F(\mathbf{s})$ takes all other countries' emissions as given and maximizes $W^f\left(e_i^f, e_a^c, e_a^f, \kappa_\varepsilon, m\right)$ with respect to $e_i^f \in \mathbb{R}_+$.

For any given membership profile $\mathbf{s} \in \Psi$, the Nash equilibrium of the stage 2 emissions game is a tuple $\left(e^{c*}, e^{f*}\right) \in \mathbb{R}_+^2$ satisfying $e_i^{c^*} = e^{c*}$ for all $i \in C(\mathbf{s})$, $e_i^{f^*} = e^{f*}$ for all $i \in F(\mathbf{s})$ and

$$\begin{aligned} \sum_{i\in C(\mathbf{s})} W^c\left(e^{c*}, e^{c*}, e^{f*}, m, \kappa_\varepsilon\right) &\geq \sum_{i\in C(\mathbf{s})} W^c\left(e^c, e^c, e_i^{f*}, m, \kappa_\varepsilon\right) \qquad \forall\, e^c \in \mathbb{R}_+, \qquad (11a) \\ W^f\left(e^{f*}, e^{c*}, e^{f*}, m, \kappa_\varepsilon\right) &\geq W^f\left(e^f, e^{c*}, e^{f*}, m, \kappa_\varepsilon\right) \qquad \forall\, e^f \in \mathbb{R}_+. \qquad (11b) \end{aligned}$$

On the assumption that the damage is linear ($D'' = 0$) the calculations in Appendix

A yield the equilibrium emissions

$$
\begin{aligned}
e_i^{c*} = \mathcal{E}^c(m, \kappa_\varepsilon) &= \frac{\alpha - [m + (n-m)\kappa_\varepsilon]\delta}{\beta} && \forall\, i \in C(\mathbf{s}), && (12)\\
e_i^{f*} = \mathcal{E}^f(m, \kappa_\varepsilon) &= \frac{\alpha - [1 + (n-1)\kappa_\varepsilon]\delta}{\beta} && \forall\, i \in F(\mathbf{s}), && (13)\\
e_i^{f*} - e_i^{c*} = \mathcal{E}^f(m, \kappa_\varepsilon) - \mathcal{E}^c(m, \kappa_\varepsilon) &= \frac{(1-\kappa_\varepsilon)(m-1)\delta}{\beta} > 0. && && (14)
\end{aligned}
$$

Inspection of (12) - (14) shows that increasing κ_ε reduces the emissions of coalition and fringe countries. In both groups the emissions decline because the probability weights (4) relating to the hypothetical material payoffs (5) of country i increase the more, the greater the number of replacements. It is unexpected, however, that the reduction of $\mathcal{E}^f$ is greater than that of $\mathcal{E}^c$. The reason is an additional effect specific to coalition countries. Increasing the replacements in the hypothetical material payoff of coalition country $j \neq i$ reduces country i's weight in country j's hypothetical material payoff. That effect is opposite to the weight shifting effect in country $i's$ hypothetical material payoff without overcompensating it. Note also that it does not necessarily follow from[6] $\left|\mathcal{E}^f_{\kappa_\varepsilon}\right| > \left|\mathcal{E}^c_{\kappa_\varepsilon}\right|$ that the reduction in moral welfare is greater in fringe countries than in coalition countries. Since $\mathcal{E}^f$ is always greater than $\mathcal{E}^c$ and the moral welfare function of each country is increasing (on the relevant domain) and strictly concave in own emissions, the reverse inequality is possible with respect to moral welfare.

According to (12) and (13), an increase in the coalition size m reduces the emissions of the coalition countries, it leaves the fringe countries' emissions unchanged $\left(\mathcal{E}^f_m = 0\right)$, and so increases the positive difference $\mathcal{E}^f(m, \kappa_\varepsilon) - \mathcal{E}^c(m, \kappa_\varepsilon)$. We infer from (14) that $\mathcal{E}^f(m, \kappa_\varepsilon) - \mathcal{E}^c(m, \kappa_\varepsilon)$ tends to zero, if m tends to $m = 1$.

4 The stage 1 membership game with ε-moral and μ-moral countries

The stage 2 emissions game analyzed above determines the countries' emissions for some given membership profile $\mathbf{s} = (s_1, \ldots, s_n) \in \Psi$. If the membership profile $\mathbf{s}$ with $m = \sum_{j=1}^n s_j$ coalition countries is given and if a country in the coalition or fringe anticipates all equilibrium emissions (12) and (13) of the stage 2 emissions game, its expected material payoff in stage 1 is equal to

$$\Pi^{h1}(m, \kappa_\varepsilon) := B\left[\mathcal{E}^h(m, \kappa_\varepsilon)\right] - D\left[m\mathcal{E}^c(m, \kappa_\varepsilon) + (n-m)\mathcal{E}^f(m, \kappa_\varepsilon)\right] \quad \text{for}\, h = c, f. \quad (15)$$

[6] Upper-case letters denote functions and subscripts attached to them indicate partial derivatives.

One possible way to proceed with the analysis is to consider the material payoff functions (15) to be countries' decision functions in stage 1. In that case, the countries are portrayed as being purely self-interested with respect to their membership decision. However, joining or not an environmental coalition is an important moral issue in its own right. So, we proceed with specifying the countries' morality with respect to membership. After that, we account for membership morality in the payoffs (15) and develop in several steps the payoff functions the countries use in the stage 1 membership game, where they act morally with respect to both emissions and membership.

We model countries of the homo moralis type with respect to membership in analogy to the countries of the homo moralis type with respect to emissions studied above. While we assumed the countries in stage 2 to act on the hypothesis that all other countries choose the same level of emissions as them with probability $\kappa_\varepsilon \in [0,1]$, we now assume that the countries act in stage 1 on the hypothesis that all other countries make the same membership decision as them with probability $\kappa_\mu \in [0,1]$. The material payoffs (15) will play the same role in the stage 1 membership game as the material payoffs (1) in the stage 2 emissions game.

The following Example 2 illustrates the calculus of a country i acting on the hypothesis that all other countries make the same membership decision as it with probability $\kappa_\mu \in [0,1]$.

Example 2. Suppose there are $n = 4$ countries (as in Example 1 above) and let the membership profile[7]. $\left(s_i^c, s_1^c, s_2^f, s_3^f\right)$ be given. If country i hypothesizes that the other countries make the same membership decision as it with probability $\kappa_\mu \in [0,1]$, it faces the hypothetical membership profiles

- $\left(s_i^c, s_1^c, s_2^f, s_3^f\right)$ with probability $1-\kappa_\mu)^3$, if $q = 0$,
- $\left(s_i^c, s_i^c, s_2^f, s_3^f\right)$, $\left(s_i^c, s_1^c, s_i^c, s_3^f\right)$, $\left(s_i^c, s_1^c, s_2^f, s_i^c\right)$ each with probability $\kappa_\mu(1-\kappa_\mu)^2$, if $q = 1$,
- $\left(s_i^c, s_i^c, s_i^c, s_3^f\right)$, $\left(s_i^c, s_i^c, s_2^f, s_i^c\right)$, $(s_i^c, s_1^c, s_i^c, s_i^c)$ each with probability $\kappa_\mu^2(1-\kappa_\mu)$, if $q = 2$,
- $(s_i^c, s_i^c, s_i^c, s_i^c)$ with probability κ_μ^3, if $q = 3$.

Consider a membership profile $\mathbf{s} \in \Psi$ in a world with $n \geq 3$ μ-moral countries, in which country i replaced the membership decisions of q from among the other $n-1$ countries by its own decision and distributed the replacements among both groups in some specific way. The payoff (15) pertaining to such a membership profile is realized with probability $\kappa_\mu^q(1-\kappa_\mu)^{n-1-q}$. Since the number of replacements $q \in \{0,1,\ldots,n-1\}$ can be attained by

[7] We denote the action s_i by s_i^f, if $s_i = 0$, and by s_i^c, if $s_i = 1$

$\binom{n-1}{q}$ different emissions profiles, the probability of a membership profile with q replacements is

$$\kappa_\mu^q(1-\kappa_\mu)^{n-1-q}\binom{n-1}{q}. \tag{16}$$

Summing the probabilities (16) over q yields $\sum_{q=0}^{n-1}\kappa_\mu^q(1-\kappa_\mu)^{n-1-q}\binom{n-1}{q}=1$.

To specify the differences in the membership profiles of Example 2 with respect to the number of coalition countries, our procedure is analogous to that in stage 2. We keep some q fixed and suppose that a country in the coalition or fringe replaces the emissions of r countries from its own group and the emissions of $q-r$ countries from the other group by its own emissions. The resulting material payoffs in stage 1 turn out to be equal to

$$\tilde{\Pi}^{h1}(r,q,m,\kappa_\varepsilon) := B\left[\mathcal{E}^h(m_h,\kappa_\varepsilon)\right] - D\left[m_h\mathcal{E}^c(m_h,\kappa_\varepsilon)+(n-m_h)\mathcal{E}^f(m_h,\kappa_\varepsilon)\right] \tag{17}$$

for $h=c,f$, where[8]

$$\begin{aligned} m_c &:= (1+q)s_i^c+(m-1-r)s_a^c+(n-m-q+r)s_a^f = m+q-r, && \text{(18a)}\\ m_f &:= (1+q)s_i^f+(n-m-1-r)s_a^f+(m-q+r)s_a^c = m-q+r. && \text{(18b)}\end{aligned}$$

Some comments on the payoffs (17) are in order. m_h, $h=c,f$, is the coalition size that results if m is given and a member of the coalition or fringe replaces by its own membership q membership decisions of other countries, r of which are replacements in its own group. For all feasible values of r and q, we get $m_c\{\geqq\}m\{\geqq\}m_f \iff q\{\geqq\}r$. So, apart from the case $q=r$, the membership morality induces coalition [fringe] countries to act as they would do in the absence of membership morality, if the true coalition were larger [smaller] than m.

The payoffs $\tilde{\Pi}^{h1}(r,q,m,\kappa_\varepsilon)$, $h=c,f$, turn out to be equal to the material payoffs (15), in which the coalition country i replaced the true coalition size m by $m_c>m$ and the fringe country i replaced m by $m_f<m$. We differentiate (15) with respect to m and conclude that the sign of the difference $\Pi^{c1}(m_c,\kappa_\varepsilon)-\Pi^{c1}(m,\kappa_\varepsilon)$ is ambiguous,[9] because country i understates both its true emissions benefits $B(\cdot)$ and the environmental damage $D(\cdot)$. If country i is in the fringe, it understates its material payoff, $\Pi^{f1}(m_f,\kappa_\varepsilon)<\Pi^{f1}(m,\kappa_\varepsilon)$, because its emissions benefits $B(\cdot)$ remain unchanged, while the damage $D(\cdot)$ increases.

[8] Analogous to e_i^c, e_i^f we write s_i^f for $s_i=0$, and s_i^c for $s_i=1$.

[9] Applying the parametric functions (3), we obtain $\Pi_m^{c1}\gtreqless 0$ if and only if $\kappa_\varepsilon \lesseqgtr \frac{m-1}{n-m}$. The proof is given in Appendix C.

Next suppose a coalition or fringe country replaced the emissions of some q other countries by its own emissions in all feasible ways. Then, its pertaining expected material payoff is equal to

$$\hat{\Pi}^{h1}\left(q, m, \kappa_{\varepsilon}\right) := \sum_{r \in R^h} A^h(r, q, m) \cdot \tilde{\Pi}^{h1}\left(r, q, m, \kappa_{\varepsilon}\right) \tag{19}$$

for $h = c, f$, where the probabilities $A^h(r, q, m)$, are the same as those defined in (8).

Finally, we recall from (16) that a country with degree of membership morality κ_μ replaces the membership of q countries with probability $\kappa_\mu^q(1-\kappa_\mu)^{n-1-q}\binom{n-1}{q}$. Accounting for all feasible values of q it follows that the material payoff a coalition or fringe country expects on the hypothesis that all other countries would choose the same membership as it with probability κ_μ is equal to

$$\begin{aligned}\mathcal{W}^h\left(m, \kappa_{\varepsilon}, \kappa_{\mu}\right) &:= \sum_{q=0}^{n-1} \kappa_\mu^q (1-\kappa_\mu)^{n-1-q} \binom{n-1}{q} \cdot \hat{\Pi}^{h1}(q, m, \kappa_\varepsilon), \\ &= \sum_{q=0}^{n-1} \kappa_\mu^q (1-\kappa_\mu)^{n-1-q} \binom{n-1}{q} \cdot \sum_{r \in R^h} A^h(r, q, m) \cdot \tilde{\Pi}^{h1}(r, q, m, \kappa_\varepsilon).(20)\end{aligned}$$

The welfare functions (20) depend on the degrees of morality κ_ε and κ_μ, which makes explicit that the full two-stage game accounts for morality on two dimensions, i.e. it accounts for morality with respect to emissions and with respect to membership. We refer to $\mathcal{W}^h$, $h = c, f$, in (20) as country i's $\varepsilon\mu$-moral welfare and assume that the countries apply these welfare functions as their decision functions in the membership game.

It is straightforward to see that a membership profile $\mathbf{s}^* \in \Psi$ with the implied coalition size $m^* = \sum_{j=1}^n s_j^*$ constitutes a Nash equilibrium of the two-stage IEA game, if

$$\begin{aligned} &S\left(m^*, \kappa_\varepsilon, \kappa_\mu\right) \geq 0 && \text{(internal stability)}, \\ &S\left(m^*+1, \kappa_\varepsilon, \kappa_\mu\right) < 0 && \text{(external stability)}, \end{aligned} \tag{21}$$

where

$$S(m, \kappa_\varepsilon, \kappa_\mu) := \mathcal{W}^c(m, \kappa_\varepsilon, \kappa_\mu) - \mathcal{W}^f(m-1, \kappa_\varepsilon, \kappa_\mu) \tag{22}$$

is known in the literature as *stability function.*

To determine stable coalitions, it is analytically convenient to treat the variable m as a non-negative real number although the coalition size is an integer. It can be shown that if a real number m' satisies $S(m', \kappa_\varepsilon, \kappa_\mu) = 0$ and $S(m'+1, \kappa_\varepsilon, \kappa_\mu) < 0$, the conditions of (weak) internal stability $S(m, \kappa_\varepsilon, \kappa_\mu) \geq 0$ and (weak) external stability $-S(m+1, \kappa_\varepsilon, \kappa_\mu) = 0$ are satisfied for all m in the unit interval $[m'-1, m']$. That interval contains either one integer, which is then the true size of a stable coalition, or both boundary points $m'-1$ and m' are integers. In the latter case we take the larger number m' to be the (unique) size of the stable coalition based on an equilibrium selection argument.

There is a function $\mathcal{M} : [0,1]^2 \to \mathbb{R}_+$ such that $m = \mathcal{M}(\kappa_\varepsilon, \kappa_\mu)$ if $S(m, \kappa_\varepsilon, \kappa_\mu) = 0$ and the condition $S(m+1, \kappa_\varepsilon, \kappa_\mu) < 0$ holds.[10,11] To avoid clumsy wording, we refer in the following to the real number $\mathcal{M}(\kappa_\varepsilon, \kappa_\mu)$ as the size of the stable coalition keeping in mind that the true size of the stable coalition is the largest integer smaller than or equal to $\mathcal{M}(\kappa_\varepsilon, \kappa_\mu)$.

Apart from the size of the stable coalition, $\mathcal{M}\,(\kappa_\varepsilon, \kappa_\mu)$, the solution of the game yields the material payoff evaluated at the size of the stable coalition

$$\mathcal{W}^h\left[\mathcal{M}(\kappa_\varepsilon, \kappa_\mu), 0, 0\right] =: \mathcal{P}^h\left(\kappa_\varepsilon, \kappa_\mu\right) \qquad \text{for } h = c, f. \tag{23}$$

That equilibrium material payoff function $\mathcal{P}^h : [0,1]^2 \to \mathbb{R}_+$ is an important piece of information, because the prevailing (consequentialist) view on mitigation policy is that the size of the stable coalition is not an end but rather a means to reduce aggregate emissions in order to enhance the countries' material payoff. Larger stable coalitions yield larger material payoffs, indeed. But as we will investigate in detail below, moral behavior impacts on material payoff not only through changing the size of the stable coalition. That is why the adequate assessment of the impact of moral behavior on mitigation requires considering the material payoff along with the coalition size.

5 The solution of the IEA game

In this section we explore the conditions for stable environmental coalitions and the associated payoffs and welfares. In order to disentangle the impact of emissions morality and membership morality we proceed in three steps. Section 5.1 analyzes the outcome of the game if countries exhibit emissions morality only ($\kappa_\varepsilon \in [0,1], \kappa_\mu = 0$). In section 5.2 we assume that the countries are moral with respect to membership only ($\kappa_\varepsilon = 0, \kappa_\mu \in [0,1]$). Section 5.3 deals with countries that exhibit both emissions and membership morality

[10] Observe that $\mathcal{M}(\kappa_\varepsilon, \kappa_\mu) = 1$ if $S(m, \kappa_\varepsilon, \kappa_\mu) < 0$ for all $m \in [2, n]$. In addition, it holds $\mathcal{M}(\kappa_\varepsilon, \kappa_\mu) = n$ if $S(m, \kappa_\varepsilon, \kappa_\mu) > 0$ for all $m \in [2, n]$.

[11] Note that $\mathcal{M}(\kappa_\varepsilon, \kappa_\mu)$ is equal to and specifies the determinants of the number m' discussed in the previous paragraph.

($\kappa_\varepsilon \in [0,1], \kappa_\mu \in [0,1]$). Section 5.4 shifts the focus from characterizing stable coalitions to the comparison of equilibrium material payoffs. All results presented in section 5 are analytically derived in Appendix with the parametric functional forms (3).[12] Some of the illustrations are based on the parameters $\alpha = 1000$, $\beta = 100$ and $\delta = 1$.

5.1 Morality with respect to emissions only ($\kappa_\varepsilon \in [0,1], \kappa_\mu = 0$)

We begin with the special case $\kappa_\varepsilon \in [0,1]$ and $\kappa_\mu = 0$ and find that[13]

$$\mathcal{M}(\kappa_\varepsilon, 0) = \begin{cases} \frac{3-(2n-1)\kappa_\varepsilon}{1-\kappa_\varepsilon} & \text{if } \kappa_\varepsilon < \frac{1}{2n-3}, \\ 1 & \text{if } \kappa_\varepsilon \geq \frac{1}{2n-3} \end{cases} \tag{24}$$

is the size of a stable coalition. It is easy to see that (24) yields $m = \mathcal{M}(0,0) = 3$, if the countries do not act morally at all. That boundary case is the standard result for purely self-interested countries in the literature (Carraro and Siniscalco 1991, Hoel 1992, Finus and Maus 2008). Here, it serves as a convenient benchmark for characterizing the outcome of morally acting countries. For all $\kappa_\varepsilon \in \left[0, \frac{1}{2n-3}\right[$ the function (24) can be shown to be strictly decreasing and strictly concave in κ_ε, and for all $\kappa_\varepsilon \in \left[\frac{1}{2n-3}, 1\right]$ it satisfies $\mathcal{M}(\kappa_\varepsilon, 0) = 1$. From these properties combined with the equivalence $m \geq 1 \iff \kappa_\varepsilon \leq \frac{1}{2n-3}$ it follows (i) that stable coalitions with $m = 2$ members form, if $\kappa_\varepsilon \in \left]0, \frac{1}{2n-3}\right]$, and (ii) that there does not exist a stable coalition if $\kappa_\varepsilon \in \left[\frac{1}{2n-3}, 1\right]$. Note also that these results hold for any number of countries.

We summarize these results in

Proposition 1. *(Coalition formation with $\kappa_\varepsilon \in [0,1]$ and $\kappa_\mu = 0$)*

(i) If $\kappa_\varepsilon = 0$, there is a stable coalition with three members.

(ii) There is a stable coalition with two members for all $\kappa_\varepsilon \in \left]0, \frac{1}{2n-3}\right[$ and there is no stable coalition for all $\kappa_\varepsilon \in \left[\frac{1}{2n-3}, 1\right]$.

The unexpected if not counterintuitive message of Proposition 1 is that increasing the degree of emissions morality generates pressure towards reducing the size of the stable coalition. That pressure decreases the membership of the stable coalition from $m = 3$ to $m = 2$ to

[12] These (or similar) parametric functions are widely used in studies of IEA games, because analytical relief is urgently needed to cope with the complexity of derivations.

[13] The derivation of (24) and the proof of Proposition 1 can be found in Appendix C.

$m = 1$. There is no stable coalition for all κ_ε greater than $1/(2n-3)$, which is a small number if n is large.[14]

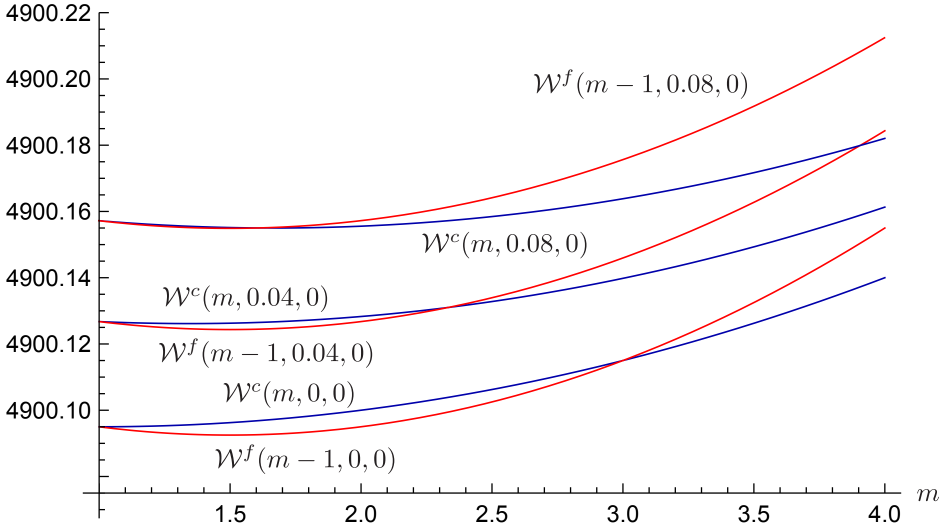

Figure 1: The dependence of the stable coalition size on κ_ε for $\alpha = 1000$, $\beta = 100$, $\delta = 1$ and $n = 10$

To explain the performance of ε-moral countries with regard to coalition formation, Figure 1 illustrates[15] the function $\mathcal{M}(\kappa_\varepsilon, 0)$ for the values $\kappa_\varepsilon = 0$, $\kappa_\varepsilon = 0.04$ and $\kappa_\varepsilon = 0.08$. Consider first the lowest pair of curves in Figure 1 that depict the welfares $\mathcal{W}^c$ and $\mathcal{W}^f$ as functions of m in the benchmark case $\kappa_\varepsilon = \kappa_\mu = 0$. The $\mathcal{W}^c$-curve is increasing whereas the $\mathcal{W}^f$-curve is u-shaped. For $m \geq 2.5$, both curves' positive slopes reflect the stronger emissions reduction of larger coalitions. The fringe countries' steeper slope is due to their freeride on the coalition countries' efforts. In Figure 1, the origin of the $\mathcal{W}^f$-curve is shifted to the right by one unit such that for any m on the m-axis the points corresponding to m on the welfare curves are $\mathcal{W}^c(m, 0, 0)$ and $\mathcal{W}^f(m-1, 0, 0)$, respectively. Taking the difference of these levels of welfare, it immediately follows from (21) that

$$\left.\begin{array}{r} S(m,0,0) \geq \\ S(m+1,0,0) < \end{array}\right\} 0 \quad \Rightarrow \quad \text{the coalition of size } m \text{ is} \left\{\begin{array}{c} \text{internally} \\ \text{externally} \end{array}\right\} \text{stable.}$$

[14] In the corresponding game in Ulph and Ulph (2023) and Eichner and Pethig (2024) emissions morality leaves unchanged the stable three-country coalition that forms in the absence of morality.

[15] With the exception of Figure 1, where we have chosen $n = 10$, in all other examples we have set $n = 100$. We selected $n = 10$ in Figure 1, because it allows a clear presentation of three welfare curves in the same diagram.

The $\mathcal{W}^f$-curve for $\kappa_\varepsilon = \kappa_\mu = 0$ intersects the $\mathcal{W}^c$-curve from below at $m = 3$ which is consistent with Proposition 1(i). Inspection of the intervals on the m-axis in which the $\mathcal{W}^f$-curve lies either above or below the $\mathcal{W}^c$-curve in Figure 1 yields that the coalition is internally stable for all $m \in [1, 3]$ and internally unstable for all $m > 3$. Of special interest is that for $m = 3$ $S(m + 1, 0, 0)$ is negative in Figure 1, because that sign of the difference is required by condition (21) for the coalition of size $m = 3$ to be externally stable. This proves that $m = 3$ is the size of the stable coalition.

Next, we assume that κ_ε is increased from $\kappa_\varepsilon = 0$ to $\kappa_\varepsilon = 0.04$. In Figure 1 the consequence is the move from the lowest pair of curves to the intermediate pair. An obvious reason for that upward shift of both curves are the properties of the equilibrium emissions in (12) and (13). Since

$$\mathcal{E}^c(m, \kappa_\varepsilon) > \mathcal{E}^f(m, \kappa_\varepsilon) \quad \text{and} \quad \mathcal{E}^f_{\kappa_\varepsilon} = -\frac{(n-1)\delta}{\beta} < \mathcal{E}^c_{\kappa_\varepsilon} = -\frac{(n-m)\delta}{\beta} < 0$$

the upward shifts do not leave the curvatures unchanged, i.e. the differences $\mathcal{W}^h(m - 1, 0.04, 0) - \mathcal{W}^h(m, 0.04, 0)$, $h = c, f$, are not the same for coalition and fringe countries and are not constant for all m. In fact, inspection of Figure 1 suggests that for $m \leq 3$ the vertical difference $\mathcal{W}^f(m - 1, \kappa_\varepsilon, 0) - \mathcal{W}^f(m - 1, 0, 0)$ is larger than the vertical difference $\mathcal{W}^c(m, \kappa_\varepsilon, 0) - \mathcal{W}^c(m, 0, 0)$ which can be confirmed analytically.[16] Due to that feature, the intersection point of the $\mathcal{W}^c$- and $\mathcal{W}^f$-curves shifts from $m = 3$ to about $m = 2.3$, if κ_ε increases from $\kappa_\varepsilon = 0$ to $\kappa_\varepsilon = 0.04$. As established above, the size of the stable coalition is the largest integer smaller than or equal to the m-coordinate of the intersection points in Figure 1. So, the move from $\kappa_\varepsilon = 0$ to $\kappa_\varepsilon = 0.04$ reduces the size of the stable coalition form $m = 3$ to $m = 2$. Economically speaking, the different way the welfare of fringe and coalition countries responds to increases in κ_ε aggravates the fringe countries' reluctance to join the coalition.

The increase of κ_ε from $\kappa_\varepsilon = 0.04$ to $\kappa_\varepsilon = 0.08$ is illustrated in Figure 1 by the move from the intermediate pair of curves to the highest pair of curves, and the assessment is analogous to the previous discussion of the increase of from $\kappa_\varepsilon = 0$ to $\kappa_\varepsilon = 0.04$. If $\kappa_\varepsilon = 0.08$, the curves $\mathcal{W}^c$ and $\mathcal{W}^f$ intersect at about $m = 1.6$ implying that there is no stable coalition.

In the absence of a stable coalition all countries act non-cooperatively which is used

[16] Follows from $\Phi^c(m) := \mathcal{W}^c(m, \kappa, 0) - \mathcal{W}^c(m, 0, 0) = \frac{(n-m)\delta^2\kappa[2(n-1)+(n-m)\kappa]}{2\beta}$, $\Phi^f(m) := \mathcal{W}^f(m - 1, \kappa, 0) - \mathcal{W}^f(m - 1, 0, 0) = \frac{\delta^2\kappa[(n^2-2n)(2-\kappa)-2(m^2-3n+1)-\kappa]}{2\beta}$, $\Phi^f(m) - \Phi^c(m) = \frac{(m-1)\delta^2\kappa[2(n-m)(1-\kappa)+2-\kappa(m-1)]}{2\beta}$ and verifying that $\Phi^f(m) - \Phi^c(m) > 0$ if $m \leq 3$.

to be referred to as business-as-usual behavior. But in contrast to that behavior of purely self-interested countries, non-cooperative ε-moral countries reduce their emissions the more, the larger is κ_ε. The consequence is, as illustrated in Figure 1, that the moral welfare of coalition and fringe countries is larger with positive κ_ε than with $\kappa_\varepsilon = 0$, and it is increasing in κ_ε, ceteris paribus. In the extreme case of $\kappa_\varepsilon = 1$, each country acts on the hypothesis that all other countries choose the same level of emissions as it. That is, each country chooses the emissions that maximize its own material payoff and symmetry implies that the maximum aggregate material payoff is attained in the absence of cooperation. The same result would be achieved if countries who do not act morally would form the grand coalition.

5.2 Morality with respect to membership only ($\kappa_\varepsilon = 0, \kappa_\mu \in [0, 1]$)

In section 5.2 we assume that all countries are purely self-interested with respect to their choice of emissions ($\kappa_\varepsilon = 0$) such that the material payoffs in (1) are their payoffs in the stage 2 emissions game. But we consider them to be of the homo moralis type with respect to membership. So, their payoffs in the stage 1 membership game are the equations (20) for the special case $\kappa_\varepsilon \equiv 0$. In Appendix E we prove

Proposition 2. *(Coalition formation with $\kappa_\varepsilon = 0$ and $\kappa_\mu \in [0, 1]$)*
For all $n \in [3, 200]$, there exists a threshold value $\tilde{\kappa}_\mu$, which depends on[17] *n, satisfies $\tilde{\kappa}_\mu \in [0, 1]$, and impacts the stability of coalitions as follows:*

(i) If $\kappa_\mu \in [0, \tilde{\kappa}_\mu]$, the size of the stable coalition $\mathcal{M}(0, \kappa_\mu)$ increases in κ_μ from $\mathcal{M}(0, 0) = 3$ to $\mathcal{M}(0, \tilde{\kappa}_\mu) = n$.

(ii) If $\kappa_\mu \in]\tilde{\kappa}_\mu, 1]$, the grand coalition is stable.

Conforming to intuition, the membership morality generates pressure towards enlarging the coalition which results in progressively increasing stable coalitions. Figure[18] 2a exhibits a free-hand graph of the function $\mathcal{M}(0, \kappa_\mu)$ with the properties established in Proposition 2. It is representative for all $n \in [3, 200]$, because the shape of the curve varies with the number of countries only slightly in two aspects: the threshold value $\tilde{\kappa}_\mu$ is increasing in n without exceeding $\tilde{\kappa}_\mu = 0.293$ and the horizontal segment of the curve moves upward as n increases. The remarkable feature is that starting at $\kappa_\mu = 0$ the size of the stable coalition is increasing steeply and progressively in κ_μ such that the grand coalition is attained at the

[17] $\tilde{\kappa}_\mu$ is strictly monotone increasing in n but bounded from above by $\bar{\kappa}_\mu = 1 - \frac{1}{\sqrt{2}} \approx 0.293$.

[18] If a figure, say Figure x, consists of two panels, we denote the figure in the left panel as Figure xa and the figure in the right panel as Figure xb.

intermediate degree of morality $\tilde{\kappa}_\mu$, and $m = n$ is sustained for all $\kappa_\mu \geq \tilde{\kappa}_\mu$. So, increasing degrees of μ-morality turn out to be very effective with respect to coalition formation.[19]

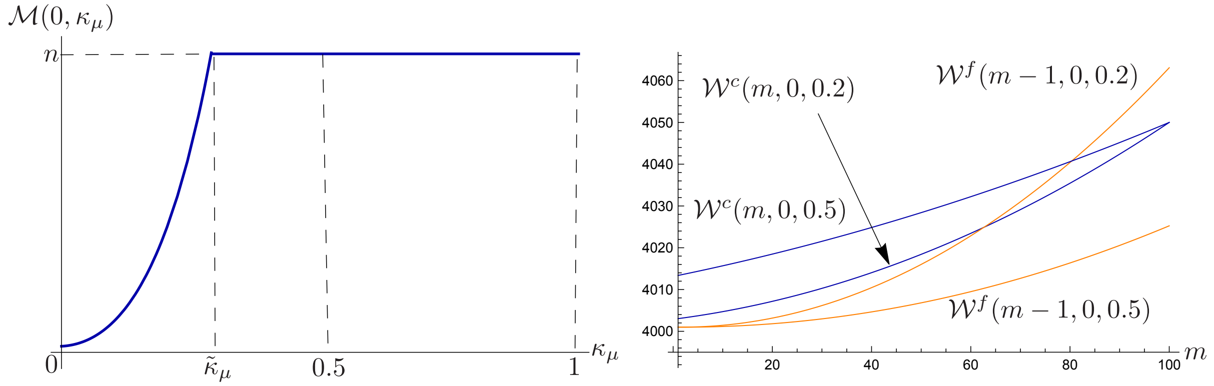

Figure 2: The size of the stable coalition depending on κ_μ and stability analysis for $\kappa_\mu = 0.2$ and $\kappa_\mu = 0.5$

Figure 2b provides insight in the transition from the upward sloping part of the $\mathcal{M}$-curve to the grand coalition by illustrating for $n = 100$ the move from $\kappa_\mu = 0.2$ to $\kappa_\mu = 0.5$. The threshold value $\tilde{\kappa}_\mu$ is approximately equal to 0.289 such that $0.2 < \tilde{\kappa}_\mu < 0.5$. Consider first the $\mathcal{W}^f$- and $\mathcal{W}^c$-curves in Figure 2b for $\kappa_\mu = 0.2$. Since the $\mathcal{W}^f$-curve intersects the $\mathcal{W}^c$-curve from below at $m = \mathcal{M}(0.2, 0)$, which we calculated as being approximately equal to $m = 62.677$, we conclude that the size of the stable coalition is $m = 62$. If we raise the degree of μ-morality from 0.2 to 0.5, the difference $\mathcal{W}^c(m, 0.5, 0) - \mathcal{W}^c(m, 0.2, 0)$ is positive (and decreasing in m) and the difference $\mathcal{W}^f(m-1, 0.5, 0) - \mathcal{W}^f(m-1, 0.2, 0)$ is negative (and increasing in m in absolute terms). The $\mathcal{W}^f$-curve shifts downward, because, as explained in section 4, μ-moral fringe countries understate the true size of the coalition, and the $\mathcal{W}^c$-curve shifts upward, because μ-moral coalition countries overstate the true size of the coalition. The shifts of the $\mathcal{W}^f$- and $\mathcal{W}^c$-curves in opposite directions are so strong that the curves corresponding to $\kappa_\mu = 0.5$ do not intersect anymore. Inspection of Figure 2b shows that in case of $\kappa_\mu = 0.5$ the coalitions of all sizes $m \in [1, 99]$ are internally stable, but externally unstable. The coalition of size $m = 100$ is also internally stable, but since there is no fringe country anymore, the grand coalition is stable.

The answer to the question, how the outcomes of the games compare when countries are either ε-moral only (section 5.1) or μ-moral only (section 5.2) is straightforward: their performance with respect to the coalition size is diametrically opposed: The ε-morality pre-

[19] Figure 2b is based on the parameters $\alpha = 1000$, $\beta = 100$, $\delta = 1$ and $n = 100$.

vents the formation of stable coalitions, whereas the μ-morality enlarges the size of stable coalitions very effectively. The performance of ε- and μ-morality with respect to the equilibrium material payoffs $\mathcal{P}^h\left(\kappa_\varepsilon,\kappa_\mu\right)$ defined in (23) can be conveniently compared in Figure 3.

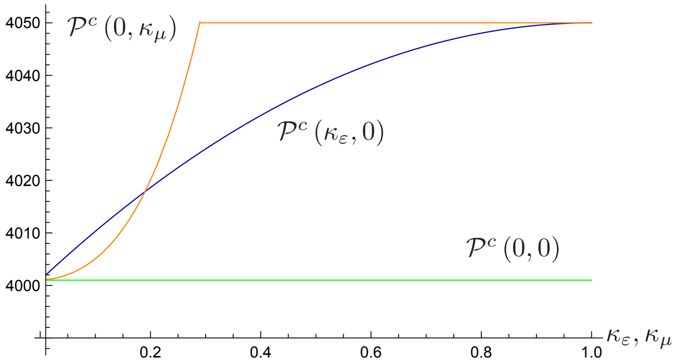

Figure 3: Equilibrium material payoffs of ε-moral or μ-moral countries for $\alpha = 1000$, $\beta = 100$, $\delta = 1$ and $n = 100$

That figure contains the graphs of the equilibrium material payoffs of coalition countries[20] $\mathcal{P}^c\left(0,0\right)$, $\mathcal{P}^c\left(\kappa_\varepsilon,0\right)$ and $\mathcal{P}^c\left(0,\kappa_\mu\right)$. It demonstrates that compared to the benchmark payoff in the absence of moral behavior, which is represented by the horizontal line in Figure 3, higher payoffs are attained if countries are either ε-moral or μ-moral. The shape of the $\mathcal{P}^c\left(0,\kappa_\mu\right)$-curve is roughly similar to the shape of the $\mathcal{M}(0,\kappa_\mu)$-curve in Figure 2a, that is, payoffs are progressively increasing in κ_μ in about the same way as stable coalitions are, and the maximum possible payoff is achieved along with the grand coalition at a rather low level of κ_μ and is maintained at all higher κ_μ. In contrast, the payoffs $\mathcal{P}^c\left(\kappa_\varepsilon,0\right)$ are increasing in κ_ε and reach the maximum payoff not before $\kappa_\varepsilon = 1$. The increase in $\mathcal{P}^c\left(\kappa_\varepsilon,0\right)$ occurs although no (large) stable coalitions are formed.

We conclude that ε- and μ-morality are substitutes with respect to material payoff if we compare the outcome of games with countries that are either ε-moral only or μ-moral only. However, equal levels of κ_ε and κ_μ yield different payoffs, in general. Denoting by $\hat{\kappa}_\varepsilon(\approx 0.2)$ in Figure 3 the intersection point of the $\mathcal{P}^c\left(\kappa_\varepsilon,0\right)$- and $\mathcal{P}^c\left(0,\kappa_\mu\right)$-curves, we find that for all $\kappa'_\varepsilon = \kappa'_\mu \in]0,\hat{\kappa}_\varepsilon[$ the material payoff is greater in games where all countries' morality is $\left(\kappa'_\varepsilon,\kappa_\mu = 0\right)$ than in games where all countries' morality is $\left(\kappa_\varepsilon = 0,\kappa'_\mu\right)$. For all

[20] It can be shown that the equilibrium material payoffs of fringe countries are larger than those of coalition countries. But as that difference in payoffs is very small, it suffices to plot the coalition countries' payoffs.

$\kappa'_\varepsilon = \kappa'_\mu \in]\hat{\kappa}_\varepsilon, 1[$ the opposite ranking holds.

The sections 5.1 and 5.2 provided useful insights in the formation of coalitions and the associated material payoffs, if countries are either ε-moral only or μ-moral only. In the subsequent section 5.3 we assume that moral countries are ε-moral as well as μ-moral.

5.3 Morality with respect to emissions and membership $((\kappa_\varepsilon, \kappa_\mu) \in [0,1] \times [0,1])$

The characterization of coalition formation in the sections 5.1 and 5.2, where countries are either ε-moral only $((\kappa_\varepsilon, \kappa_\mu) \in [0,1] \times \{0\})$ or μ-moral only $((\kappa_\varepsilon, \kappa_\mu) \in \{0\} \times [0,1])$, is theoretically interesting in its own right. That countries exhibit moderate Kantian behavior either with respect to emissions only or with respect to membership only is possible, but serious moralists are likely to behave morally 'on both dimensions'. Their moral stance may be a weakly asymmetric morality ($\kappa_\varepsilon > 0$, $\kappa_\mu > 0$, $\kappa_\varepsilon \neq \kappa_\mu$) or even the fully symmetric morality ($\kappa_\varepsilon = \kappa_\mu = \kappa$). Since the focus of the present paper is analytical rather than empirical, we will investigate coalition formation and payoffs for all feasible tuples $(\kappa_\varepsilon, \kappa_\mu)$.[21]

Proposition 3. *(Coalition formation with $\kappa_\varepsilon \in [0,1]$ and $\kappa_\mu \in [0,1]$)*
For all $n \in [9, 200]$ the following holds:

(i) There is a function $K^\varepsilon : [0,1[\to]0,1]$ such that $\mathcal{M}(\kappa_\varepsilon, \kappa_\mu)$ is weakly declining in κ_ε on the interval $[0, K^\varepsilon(\kappa_\mu)]$ and $\mathcal{M}(\kappa_\varepsilon, \kappa_\mu) = 1$ for all $\kappa_\varepsilon \geq K^\varepsilon(\kappa_\mu)$. K^ε is strictly increasing in κ_μ and has the property $\lim_{\kappa_\mu \to 1} K^\varepsilon(\kappa_\mu) = 1$.

(ii) There is a function $K^\mu : [0,1[\to]0,1]$ such that $\mathcal{M}(\kappa_\varepsilon, \kappa_\mu)$ is weakly increasing in κ_μ on the interval $[0, K^\mu(\kappa_\varepsilon)]$ and $M(\kappa_\varepsilon, \kappa_\mu) = n$ for all $\kappa_\mu \geq K^\mu(\kappa_\varepsilon)$. K^μ is strictly increasing in κ_ε and has the property $\lim_{\kappa_\varepsilon \to 1} K^\mu(\kappa_\varepsilon) = 1$.

(iii) There exist threshold values $\underset{\sim}{\kappa}$ and $\tilde{\kappa}$ with $0 < \underset{\sim}{\kappa} < \tilde{\kappa} < 0.5$ such that $\mathcal{M}(\kappa_\varepsilon, \kappa_\mu)$ is weakly increasing in $\kappa_\varepsilon = \kappa_\mu = \kappa$ on the interval $\left[\underset{\sim}{\kappa}, \tilde{\kappa}\right]$ and $\mathcal{M}(\kappa_\varepsilon, \kappa_\mu) = n$ for all $\kappa \geq \tilde{\kappa}$.

In the following, we illustrate and add some interesting details to the results of Propo-

[21] The proof of Proposition 3(iii) can be found in Appendix F.

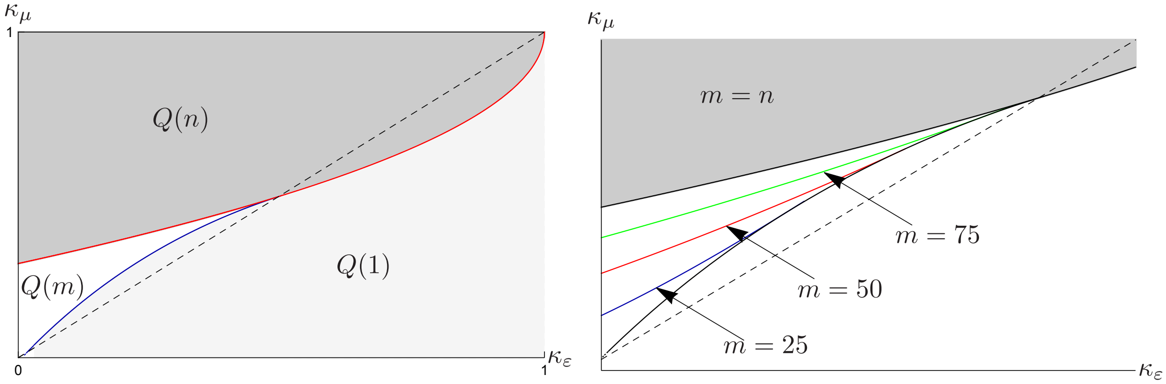

Figure 4: Isoquants for $n = 100$

sition 3 for $n = 100$.[22] The rectangular in Figure 4a represents the set $[0,1] \times [0,1]$ of feasible tuples $(\kappa_\varepsilon, \kappa_\mu)$ that is partitioned into three subsets denoted $Q(1)$, $Q(m)$ and $Q(n)$.

- $Q(1)$ is the set of $(\kappa_\varepsilon, \kappa_\mu)$-tuples that yield no coalition; it corresponds to the light grey shaded area in Figure 4a.
- $Q(m)$ is the set of $(\kappa_\varepsilon, \kappa_\mu)$-tuples that produce coalitions with at least 2 and at most 99 members; it corresponds to the small wedge-shaped unshaded area in Figure 4a.
- $Q(n)$ is the set of $(\kappa_\varepsilon, \kappa_\mu)$-tuples that yield the grand coalition; it corresponds to the dark grey area in Figure 4a.

Figure 4b exhibits an enlarged section of Figure 4a containing the set $Q(m)$. The upward sloping curves in $Q(m)$ are isoquants, i.e. curves where the coalition size is constant at the level $m = 25$, $m = 50$ and $m = 75$, respectively. If we choose a point $(\kappa_\varepsilon, \kappa_\mu)$ on such an isoquants and increase κ_ε $[\kappa_\mu]$ only, the size of the stable coalition decreases [increases]. In order to stay on the initial isoquant after increasing $(\kappa_\varepsilon, \kappa_\mu)$ by positive increments $\Delta\kappa_\varepsilon$ and $\Delta\kappa_\mu$, these increments must be chosen such that the increase in the downsizing pressure generated by $\Delta\kappa_\varepsilon$ is exactly neutralized by the increase in the upsizing pressure of $\Delta\kappa_\mu$. The principal message is that in the set $Q(m)$ ε-morality and μ-morality are no substitutes with respect to coalition formation.

A move from some tuple $(\kappa_\varepsilon, \kappa_\mu)$ to another tuple $(\kappa'_\varepsilon, \kappa'_\mu)$ in $Q(n)$ leaves the size of the coalition unchanged, because in all $(\kappa_\varepsilon, \kappa_\mu) \in Q(n)$ the upsizing pressure of κ_μ overcompensates the downsizing pressure of κ_ε even if κ_ε is very high. This is why the shifts from κ_ε to κ'_ε and/or from κ_μ to κ'_μ change the downsizing and upsizing pressures, but fail to actually downsize the initial grand coalition. We apply the same arguments to the set $Q(1)$

[22] Figure 4 graphically proves Proposition 3(i) and (ii) for $n = 100$. Analogous figures are obtained for $n \in [10, 200]$.

and conclude that moves from some tuple $(\kappa_\varepsilon, \kappa_\mu)$ to another tuple $(\kappa'_\varepsilon, \kappa'_\mu)$ in $Q(1)$ fail to generate a coalition, because in all $(\kappa_\varepsilon, \kappa_\mu) \in Q(1)$ the downsizing pressure of κ_ε overcompensates the upsizing pressure of κ_μ. In sum, we identified cases where small changes in the pressure exerted by changes in the degree of ε- or μ-morality result in either small changes or no changes at all of the coalition size.

The Propositions 3(i) and 3(ii) characterize how stable coalitions respond to partial variations of κ_ε and κ_μ, respectively. To obtain further insight into that relation, it is convenient to consider the rectangular in Figure 4a, where these variations correspond to moving along a horizontal or vertical line. In order to understand, how the impact of the partial variation of one morality parameter depends on the level at which the other is kept constant, we investigate by means of Figure 5 how the coalition size $m = \mathcal{M}(\bar{\kappa}_\varepsilon, \kappa_\mu)$ varies with κ_μ while keeping $\kappa_\varepsilon = \bar{\kappa}_\varepsilon$ fixed at a low, an intermediate and a high level. After that we will conduct the analogous analysis to partial variations of κ_μ by means of Figure 6a. Table 1 serves to disentangle the graphs in the Figures 5a and 6a.

Graph of $\mathcal{M}(\kappa_\varepsilon, \bar{\kappa}_\mu)$ in Figure 5a			Graph of $\mathcal{M}(\bar{\kappa}_\varepsilon, \kappa_\mu)$ in Figure 6a		
$\bar{\kappa}_\mu$ low	$\bar{\kappa}_\mu$ medium	$\bar{\kappa}_\mu$ high	$\bar{\kappa}_\varepsilon$ low	$\bar{\kappa}_\varepsilon$ medium	$\bar{\kappa}_\varepsilon$ high
$0AA'BE$	$0FF'CE$	$0GDE$	$DEE'H$	$ABFF'H$	$ACGH$

Table 1: Impact of κ_ε (Figure 5a) and of κ_μ (Figure 6a)] on the size of stable coalitions, when alternative levels $\bar{\kappa}_\mu$ and $\bar{\kappa}_\varepsilon$, respectively, are given for $n = 100$

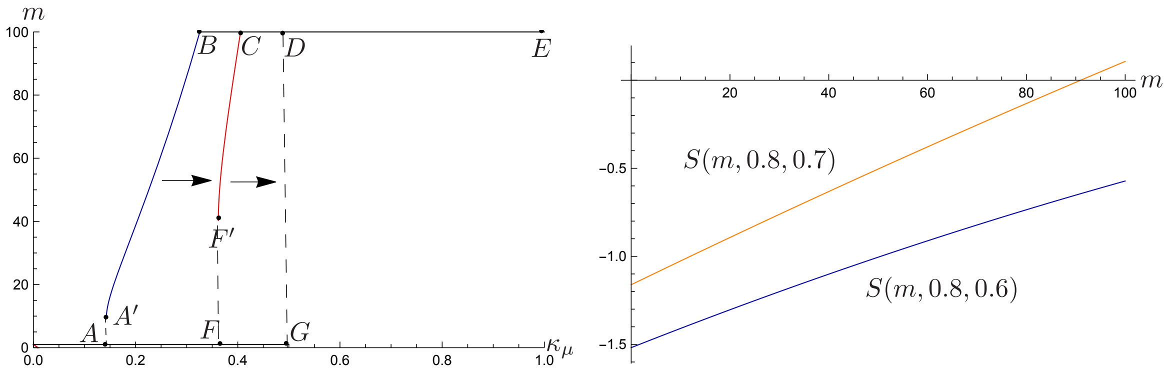

Figure 5: Stable coalitions depending on κ_μ for alternatively given $\bar{\kappa}_\varepsilon$ for $n = 100$

Consider first the case of partial variations of κ_μ. Figure 5a depicts the graph of the function $\mathcal{M}$ that results, if κ_μ increases successively from zero to one with κ_ε being fixed alternatively at some low, intermediate or high level. The assignment of the exact graph to the level of κ_ε is given in Table 1. In each case there is an interval $I(1, \bar{\kappa}_\varepsilon)$ of κ_μ which satisfies the equation $m = \mathcal{M}(\bar{\kappa}_\varepsilon, \kappa_\mu) = 1$. The interval $I(1, \bar{\kappa}_\varepsilon)$ is a proper subset of the interval $[0, 1]$ with lower bound $\kappa_\mu = 0$.

Figure 5a shows that for any given $\bar{\kappa}_\varepsilon$, increasing κ_μ weakly increases the coalition size, because the downsizing pressure of $\bar{\kappa}_\varepsilon$ is given, whereas the upsizing pressure increases along with κ_μ until eventually the latter dominates the former. Since the downsizing pressure is increasing in κ_ε, it follows that the higher $\bar{\kappa}_\varepsilon$, the larger the increase in κ_μ necessary to enlarge the coalition size. It is interesting to describe - and identify in Figure 5a - that feature from a different perspective. For every $\bar{\kappa}_\varepsilon$ there is an interval with lower bound $\kappa_\mu = 0$ such that $m = \mathcal{M}(\bar{\kappa}_\varepsilon, \kappa_\mu) = 1$ for all $\kappa_\mu \in I(1, \bar{\kappa}_\varepsilon)$, where $\partial I(1, \bar{\kappa}_\varepsilon)/\partial \bar{\kappa}_\varepsilon > 0$ and $\lim_{\bar{\kappa}_\varepsilon \to 1} I(1, \bar{\kappa}_\varepsilon) = [0, 1]$. So, the downsizing pressure of $\bar{\kappa}_\varepsilon$ dominates the upsizing pressure of all $\kappa_\mu \in I(1, \bar{\kappa}_\varepsilon)$.

Another noteworthy property of all graphs in Figure 5a is a discontinuity that increases in size when moving from the low to the high level of $\bar{\kappa}_\varepsilon$. To understand that jump in coalition size, recall first that $m = \mathcal{M}(\bar{\kappa}_\varepsilon, \kappa_\mu) = 1$ for all $\kappa_\mu \in I(1, \bar{\kappa}_\varepsilon)$ applies to all three graphs in Figure 4a. For low and intermediate levels of $\bar{\kappa}_\varepsilon$, increasing κ_μ corresponds to a move from the set $Q(1)$ to the set $Q(m)$ and from there to the set $Q(n)$. The jump occurs from $m = 1$ to $m \in [2, n[$ at the transition from $Q(1)$ to $Q(m)$, and it is increasing in $\bar{\kappa}_\varepsilon$. If $\bar{\kappa}_\varepsilon$ is sufficiently high, the coalition size jumps up from $m = 1$ to $m = n$, when crossing the line that separates $Q(1)$ from $Q(n)$. To explain these discontinuities, we consider a high $\bar{\kappa}_\varepsilon$ as an example. In that case it holds for all $\kappa_\mu \in I(1, \bar{\kappa}_\varepsilon)$ that the stability function satisfies $S(m, \bar{\kappa}_\varepsilon, \kappa_\mu) < 0$ for all $m \in [2, n]$. That is shown in Figure 5b for the tuple $(\bar{\kappa}_\varepsilon, \kappa_\mu) = (0.8, 0.6)$ that satisfies $\kappa_\mu \in I(1, \bar{\kappa}_\varepsilon)$. The graph of the corresponding stability function $S(m, 0.8, 0.6)$ is in the negative quadrant of Figure 5b. If we increase κ_μ from 0.6 to 0.7, we move from $(0.8, 0.6) \in Q(1)$ to $(0.8, 0.7) \in Q(n)$, and, as Figure 5b shows, this move shifts the graph of the stability function upward so strongly that $S(m, 0.8, 0.7)$ has become positive for $m = n = 100$. That implies $\mathcal{M}(0.8, 0.7) = n$.

In analogy to Figure 5a, Figure 6a shows how the coalition size responds to changes in κ_ε, while κ_μ is kept constant at three alternative levels. The assignment of the exact graph to the level of κ_μ is given in Table 1. The outcome is mirror symmetric to that in Figure

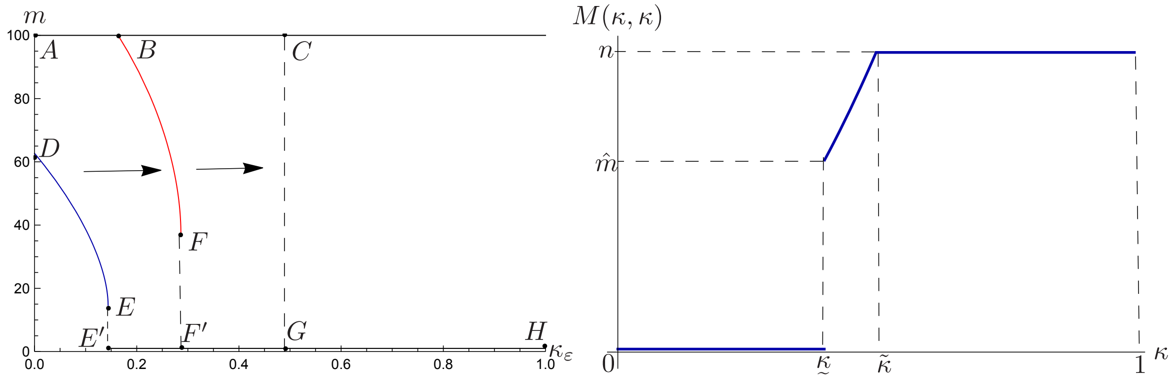

Figure 6: Stable coalitions depending on κ_ε for alternatively given $\bar{\kappa}_\mu$ for $n = 100$ (Figure 6a) and the shape of the $M(\kappa, \kappa)$-curve for $n \in]9, 200]$ (Figure 6b)

5a. For each given $\bar{\kappa}_\mu$, increasing κ_ε weakly reduces the size of the stable coalition, because the upsizing pressure of $\bar{\kappa}_\mu$ is given, whereas the downsizing pressure increases along with κ_ε until eventually the latter dominates the former. Since the upsizing pressure of κ_μ is also increasing in κ_μ, it follows that the higher $\bar{\kappa}_\mu$, the stronger the increase in κ_ε necessary to reduce the coalition size. The explanation of the discontinuities in the graphs of Figure 6a is analogous to that in Figure 5a.

Finally, some comments are in order on Proposition 3(iii) that deals with the case of symmetric morality formalized by the constraint $\kappa_\varepsilon = \kappa_\mu = \kappa$. The graph[23] in Figure 6b is the locus of the coalition size $\mathcal{M}(\kappa, \kappa)$ for all $\kappa \in [0, 1]$. Increasing κ successively from zero to one corresponds to moving along the diagonal from the lower left corner of the rectangular to the upper right corner in Figure 4a. For all low levels of $\kappa \in [0, \underset{\sim}{\kappa}]$ we get $(\kappa, \kappa) \in Q(1)$, for all intermediate levels of $\kappa \in [\underset{\sim}{\kappa}, \tilde{\kappa}]$ we get $(\kappa, \kappa) \in Q(m)$ and for all high levels of $\kappa \in [\tilde{\kappa}, 1]$ we get $(\kappa, \kappa) \in Q(n)$. The discontinuity occurs at the transition from $Q(1)$ to $Q(m)$.[24] Thus, increasing κ weakly increases the size of the stable coalition, as does increasing κ_μ, if $\bar{\kappa}_\varepsilon$ is given. To put it more pointedly, if κ is small, i.e. if $\kappa \in [0, \underset{\sim}{\kappa}]$, no coalition forms exactly as in the case where $\kappa_\varepsilon \in [0, \underset{\sim}{\kappa}]$ and $\kappa_\mu = 0$. Conversely, if κ is large, i.e. if $\kappa \in [\tilde{\kappa}, 1]$, the grand coalition forms exactly as in the case where $\kappa_\mu \in [\tilde{\kappa}, 1]$ and $\kappa_\varepsilon = 0$. We conclude that if κ is low, the downsizing pressure of κ_ε overcompensates the upsizing pressure of κ_μ, and the opposite holds, if κ is high.

[23] That figure is a free-hand drawing that slightly deviates from the numerical versions for two reasons. First, the exact interval $[\underset{\sim}{\kappa}, \tilde{\kappa}]$ is so small that one could hardly identify the upward sloping graph in that interval. Second, there is an interval of κ with lower bound zero with stable coalitions of three or two countries (similar as in Proposition 1(ii)), which we neglect in Figure 5b, because the interval is very small.

[24] In case of $n = 100$ the stable coalition jumps from $m = 1$ to $m = 96$ at $\underset{\sim}{\kappa} \approx 0.4947$. Thus, for $\kappa_\varepsilon = \kappa_\mu = \kappa$ the set $Q(m)$ is very small and not visible in Figure 4a.

The preceding analysis showed that the impact of ε- and μ-morality on the stability of coalitions is a result of complex interactions of the degrees of ε- and μ-morality. They turn out to be no substitutes with respect to the size of stable coalitions, because keeping the coalition size constant after changing one of the morality parameters does not change the other morality parameter in the opposite direction.[25] Nor are the ε- and μ-moralities complements in the sense that an increase of the coalition size is infeasible unless both morality parameters are increased simultaneously.

Next we consider the material payoff function $p = \mathcal{P}^c$ on its domain, the set of feasible tuples $(\kappa_\varepsilon, \kappa_\mu)$. Obviously, on the entire set $Q(n)$ in Figure 4a the socially optimal material payoff is attained. The set $Q(1)$ exhibits material-payoff isoquants in the form of vertical straight lines with levels of material payoff that are strictly increasing in κ_ε from the level as in BAU at $\kappa_\varepsilon = 0$ to the first best level attained at $\kappa_\varepsilon = 1$. As displayed in Figure 7a, there are upward sloping material-payoff isoquants in the set $Q(m)$ which extend into the set $Q(1)$ in the form of vertical lines with a discontinuity at the boundary line between the sets $Q(m)$ and $Q(1)$. We conclude that the ε- and μ-moralities are neither substitutes nor complements with respect to material payoff.

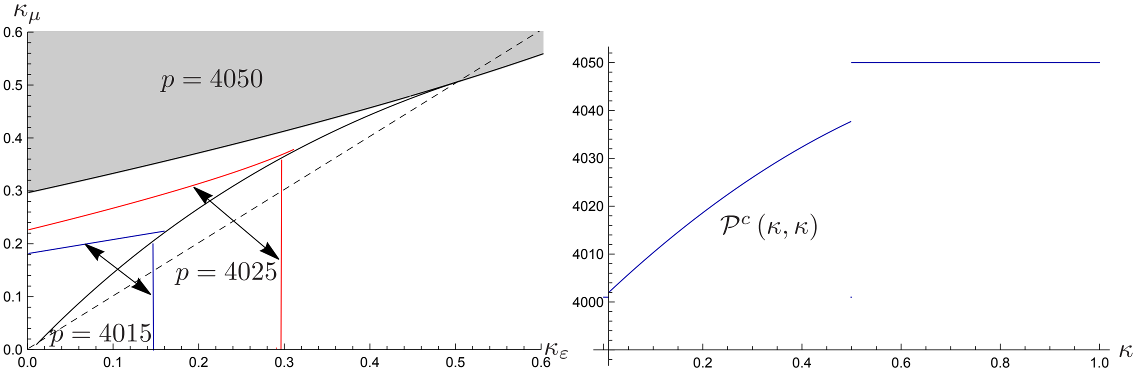

Figure 7: Material-payoff isoquants (Figure 7a) and material payoffs of symmetrically moral countries (Figure 7b) for $\alpha = 1000$, $\beta = 100$, $\delta = 1$ and $n = 100$

Analogous to the procedure we applied to the coalition size $\mathcal{M}(\cdot)$ one can investigate in detail how the material payoff $\mathcal{P}^c$ changes in response to the variation of one degree of morality while keeping the other fixed at some positive level. We refrain from performing these exercises, however, because it is easy to derive the principal curvature of $\mathcal{P}^c$ in these

[25] In geometric terms, κ_ε and κ_μ are no substitutes, because there are no negatively sloped coalition-size isoquants.

exercises from the properties of the function $\mathcal{P}^c$ established in the last paragraph. The rough overall message is that an increase in either morality parameter weakly increases the equilibrium material payoff.[26]

The equilibrium material payoffs in the case of symmetric moralities ($\kappa_\varepsilon = \kappa_\mu = \kappa$) are plotted in Figure 7b. For κ in the interval from zero up to the discontinuity, the shape of the curve $\mathcal{P}^c(\kappa, \kappa)$ is the same as that of the curve $\mathcal{P}^c(\kappa_\varepsilon, 0)$ in the corresponding interval of κ_ε in Figure 3. Furthermore, the horizontal line $\mathcal{P}^c(\kappa, \kappa)$ for high levels of κ in Figure 7b corresponds to the horizontal line $\mathcal{P}^c(0, \kappa_\mu)$ for high levels of κ_μ in Figure 3. That remarkable feature of symmetric moralities is consistent with our observation above that the countries' ε-morality dominates their μ-morality if the degree of both moralities is low and vice versa if it is high.

6 Concluding remarks

If countries act purely self-interested with regard to their decisions on both membership and emissions, as assumed in the basic IEA game of the literature, coalitions are small, emissions are excessive and material payoffs are suboptimally small. But there is empirical evidence that individuals restrain their self-interest (to some extent) in an effort 'to do the right thing', and as far as voters influence their governments' policy, moral behavior can also be expected at the country level. We found that the impact of morality on coalition size and material payoff is as expected, if countries are moral with respect to membership only. If $\kappa_\varepsilon = 0$ and κ_μ increases from zero to one, the coalition size and the material payoff strictly increase first and then stay at their maximum levels for all κ_μ larger than some rather small threshold value. If $\kappa_\mu = 0$ and κ_ε increases from zero to one, the small stable coalition of size $m = 3$ that forms if $\kappa_\varepsilon = 0$, becomes unstable and there is no stable coalition anymore for all κ_ε exceeding a very small threshold value of κ_ε. However, despite the absence of stable coalitions, the material payoff is strictly increasing in κ_ε. These findings turn out to be no good guide to predict the outcome in the presumably more relevant scenario of countries who are moral on both dimensions.

To see that, we briefly summarize the pertinent results of our analysis. Since high degrees of morality appear to be empirically irrelevant, we restrict the focus to low levels of κ_μ and κ_ε.

[26] That statement is rough because there are small segments of decreasing payoffs.

(i) If[27] $(\kappa_\varepsilon, \kappa_\mu) \in \text{int } Q(m)$, small increases in κ_ε reduce (!), and small increases in κ_μ raise both the coalition size and the material payoff. Almost all $(\kappa_\varepsilon, \kappa_\mu) \in Q(m)$ satisfy $\kappa_\mu/\kappa_\varepsilon > 1$.

(ii) If $(\kappa_\varepsilon, \kappa_\mu) \in \text{int } Q(1)$, small changes in κ_μ fail to induce a stable coalition and therefore have no impact on material payoff; small increases in κ_ε do not form a coalition either, but raise the material payoff. Most $(\kappa_\varepsilon, \kappa_\mu) \in Q(1)$ satisfy $\kappa_\mu/\kappa_\varepsilon \leq 1$. If the ratio $\kappa_\mu/\kappa_\varepsilon$ is constant and κ_ε increases (along with κ_μ), the impact on coalition formation and material payoff is the same as in the scenario $(\kappa_\varepsilon > 0, \kappa_\mu \equiv 0)$.

Obviously, the degrees of morality cannot easily be fixed and changed like conventional policy instruments. But if policy makers have some lever or nudge at their disposal for small changes of the morality parameters, the preceding considerations suggest some unexpected recommendations. If $(\kappa_\varepsilon, \kappa_\mu) \in \text{int } Q(m), \kappa_\varepsilon$ should be reduced (!) and/or κ_μ should be increased. In contrast, if $(\kappa_\varepsilon, \kappa_\mu) \in \text{int } Q(1), \kappa_\varepsilon$ should be increased, while changes in κ_μ have no effects at all. We conclude that the importance of the formation of IEAs for effective mitigation in a world of countries of the homo moralis type depends on whether the tuple $(\kappa_\varepsilon, \kappa_\mu)$ is in $Q(m)$ or in $Q(1)$.

We are aware that it is unclear how robust the results of our IEA game are because the assumptions we needed to apply for reasons of tractability are restrictive. With this caveat in mind, we interpret the global real-world climate policy in the light of our theoretical approach in which countries are of the homo moralis type with low degrees of membership and emissions morality. Since the Paris Agreement relies on nationally determined contributions, it cannot be considered as a climate coalition as modelled in our game, because that kind of coalition requires its members to commit to that level of national emissions (or contributions) which secures the maximum aggregate payoff for all members. So, the current state of climate change may be characterized (in the light of our game) by rather small positive values of κ_ε and κ_μ. But in that scenario our theory suggests that small increases in κ_μ neither promote the formation of a coalition in the theoretical sense nor have they any impact on emissions and material welfare. The only way to improve upon mitigation and material payoff is to increase κ_ε.

[27] The sets $Q(m)$ and $Q(1)$ are defined in the context of Figure 4.

Appendix

A: Derivation of the equilibrium emissions:

The fringe: Fringe country $i \in F(\mathbf{s})$ maximizes $W^f(e^f_i, e^c_a, e^f_a, m)$ with respect to e^f_i which yields the first-order condition

$$W^f_{e^f_i}(e^f_i, e^c_a, e^f_a, m, \kappa_\varepsilon) = \sum_{q=0}^{n-1} \kappa_\varepsilon^q (1-\kappa_\varepsilon)^{n-1-q} \binom{n-1}{q} \sum_{r\in R^f} A^f \cdot \left\{ B'\left(e^f_i\right) - (q+1)D'\left(\cdot\right) \right\}. \qquad \text{(A1)}$$

For linear damage function $D' = \delta$ and accounting for[28]

$$\sum_{r\in R^f} A^f = 1, \quad \sum_{r\in R^f} qA^f = q \qquad \text{(A2)}$$

the first-order condition (A1) simplifies to

$$\begin{aligned} W^f_{e^f_i}(e^f_i, e^c_a, e^f_a, m, \kappa_\varepsilon) &= B'(e^f_i) - \delta \sum_{q=0}^{n-1} (1+q) \underbrace{\kappa_\varepsilon^q (1-\kappa_\varepsilon)^{n-1-q} \binom{n-1}{q}}_{\equiv f(q, n-1, \kappa_\varepsilon)} \\ &= B'(e^f_i) - \delta \mathbf{E}\left[X+1\right] = 0, \end{aligned} \qquad \text{(A3)}$$

where $f(q, n-1, \kappa_\varepsilon)$ is the density function of a random variable X that follows the binomial distribution with parameter $n-1$ and probability $p = \kappa_\varepsilon$. Inserting $\mathbf{E}\left[X\right] = (n-1)\kappa_\varepsilon$ we get

$$W^f_{e^f_i}(e^f_i, e^c_a, e^f_a, m, \kappa_\varepsilon) = B'(e^f_i) - \left[1 + (n-1)\kappa_\varepsilon\right]\delta = 0. \qquad \text{(A4)}$$

Making use of $B(e_i) = \alpha e_i - \frac{\beta}{2}\left(e_i\right)^2$ we obtain from (A4)

$$e^f_i = \frac{\alpha - \left[1 + (n-1)\kappa_\varepsilon\right]\delta}{\beta} = \mathcal{E}^f(m, \kappa_\varepsilon). \qquad \text{(A5)}$$

The coalition: The manager of the coalition maximizes

$$\sum_{j\in C(\mathbf{s})} W^c(e^c_j, e^c_a, e^f_a, m, \kappa_\varepsilon) \qquad \text{(A6)}$$

with respect to e^c_i which yields the first-order condition

$$\sum_{j\in C(\mathbf{s})} \frac{\mathrm{d}W^c_{e^c_i}(e^c_i, e^c_a, e^f_a, m, \kappa_\varepsilon)}{\mathrm{d}e^c_i} = W^c_{e^c_i}(e^c_i, e^c_a, e^f_a, m, \kappa_\varepsilon) + (m-1)W^c_{e^c_i}(e^c_j, e^c_a, e^f_a, m, \kappa_\varepsilon), \qquad \text{(A7)}$$

[28] The first identity in (A2) is known as Vandermonde's identity.

where

$$W^c_{e^c_i}(e^c_i, e^c_a, e^f_a, m, \kappa_\varepsilon) = \sum_{q=0}^{n-1} \kappa_\varepsilon^q (1-\kappa_\varepsilon)^{n-1-q} \binom{n-1}{q} \sum_{r \in R^c} A^c \cdot \left\{ B'\left(e^c_i\right) - D'\left(\cdot\right) \right\}. \tag{A8}$$

For linear damage function $D' = \delta$ and accounting for

$$\sum_{r \in R^c} A^c = 1, \quad \sum_{r \in R^c} qA^c = q \tag{A9}$$

(A8) simplifies to

$$W^c_{e^c_i}(e^c_i, e^c_a, e^f_a, m, \kappa_\varepsilon) = B'\left(e^c_i\right) - \delta \sum_{q=0}^{n-1} (1+q) \kappa_\varepsilon^q (1-\kappa_\varepsilon)^{n-1-q} \binom{n-1}{q}. \tag{A10}$$

We can differentiate $W^c(e^c_i, e^c_a, e^f_a, m)$ with respect to any other country's emissions e^c_a or e^f_a only if these emissions are not (yet) replaced by e^c_i. Since the average non-replacement rate is equal to $\left(1 - \frac{q}{n-1}\right)$ for all countries other than i we obtain

$$W^c_{e^c_j}(e^c_i, e^c_a, e^f_a, m, \kappa_\varepsilon) = -\delta \sum_{q=0}^{n-1} \kappa_\varepsilon^q (1-\kappa_\varepsilon)^{n-1-q} \binom{n-1}{q} \left(1 - \frac{q}{n-1}\right). \tag{A11}$$

Inserting (A10) and (A11) in (A7) we get

$$\begin{aligned} \sum_{j \in C(\mathbf{s})} \frac{\mathrm{d}W^c_{e^c_i}(e^c_j, e^c_a, e^f_a, m, \kappa_\varepsilon)}{\mathrm{d}e^c_i} &= B'\left(e^c_i\right) - \delta \sum_{q=0}^{n-1} \kappa_\varepsilon^q (1-\kappa_\varepsilon)^{n-1-q} \binom{n-1}{q} \left(m + \frac{n-m}{n-1} q\right) \\ &= B'(e^c_i) - \delta \cdot \mathbf{E}\left[m + \frac{n-m}{n-1} X\right] = 0. \end{aligned} \tag{A12}$$

Accounting for $\mathbf{E}\left[X\right] = (n-1)\kappa_\varepsilon$ in (A12) we obtain

$$\sum_{j \in C(\mathbf{s})} \frac{\mathrm{d}W^c_{e^c_i}(e^c_j, e^c_a, e^f_a, m, \kappa_\varepsilon)}{\mathrm{d}e^c_i} = B'(e^c_i) - \delta \left[m + (n-m)\kappa_\varepsilon\right] = 0. \tag{A13}$$

Finally, making use of $B(e_i) = \alpha e_i - \frac{\beta}{2}\left(e_i\right)^2$ we get

$$e^c_i = \frac{\alpha - \left[m + (n-m)\kappa_\varepsilon\right]\delta}{\beta} = \mathcal{E}^c(m, \kappa_\varepsilon). \tag{A14}$$

Inserting $= \mathcal{E}^f(m, \kappa_\varepsilon)$ from (A5) and $\mathcal{E}^c(m, \kappa_\varepsilon)$ from (A14) into the material payoff function (1) yields

$$\Pi^{1f}\left(m,\kappa_\varepsilon\right) = \frac{\alpha^2 - 2\alpha n\delta + \delta^2\left[(2n-1)(1-\kappa_\varepsilon)^2 + 2(m^2-m)(1-\kappa_\varepsilon) + n^2(2-\kappa_\varepsilon)\kappa_\varepsilon\right]}{2\beta}, \tag{A15}$$

$$\Pi^{1c}\left(m,\kappa_\varepsilon\right) = \frac{\alpha^2 - 2\alpha n\delta - \delta^2\left[2m(1-\kappa_\varepsilon)(1+n\kappa_\varepsilon) - m^2(1-\kappa_\varepsilon^2) + n(n\kappa_\varepsilon^2 - 2(n-1)\kappa_\varepsilon - 2)\right]}{2\beta}. \tag{A16}$$

B: Derivation of the expected material payoffs in stage 1:

Rewriting the material payoff (A15) and (A16) as

$$\Pi^{1f}\left(\sum_{j\in N} s_j, \kappa_\varepsilon\right) = \frac{\alpha^2 - 2\alpha n\delta + \delta^2\left[(2n-1)(1-\kappa_\varepsilon)^2 + 2\left[\left(\sum_{j\in N} s_j\right)^2 - \sum_{j\in N} s_j\right](1-\kappa_\varepsilon) + n^2(2-\kappa_\varepsilon)\kappa_\varepsilon\right]}{2\beta}, \tag{A17}$$

$$\Pi^{1c}\left(\sum_{j\in N} s_j, \kappa_\varepsilon\right) = \frac{\alpha^2 - 2\alpha n\delta - \delta^2\left[2\sum_{j\in N} s_j(1-\kappa_\varepsilon)(1+n\kappa_\varepsilon) - \left(\sum_{j\in N} s_j\right)^2(1-\kappa_\varepsilon^2) + n(n\kappa_\varepsilon^2 - 2(n-1)\kappa_\varepsilon - 2)\right]}{2\beta}, \tag{A18}$$

defining

$$m_c := (1+q)\underbrace{s_i^f}_{=0} + (n-m-1-r)\underbrace{s_a^f}_{=0} + (m-q+r)\underbrace{s_a^c}_{=1} = m+q+r =: M^c(r,q,m), \tag{A19}$$

$$m_f := (1+q)\underbrace{s_i^c}_{=1} + (m-1-r)\underbrace{s_a^c}_{=1} + (n-m-q+r)\underbrace{s_a^f}_{=0} = m-q+r =: M^f(r,q,m) \tag{A20}$$

and accounting for

$$\tilde{\Pi}^{f1}(r,q,m,\kappa_\varepsilon) = \Pi^{f1}(m_f,\kappa_\varepsilon), \quad \tilde{\Pi}^{c1}(r,q,m,\kappa_\varepsilon) = \Pi^{c1}(m_c,\kappa_\varepsilon), \tag{A21}$$

the expected material payoff in stage 1 is

$$\mathcal{W}^f\left(m,\kappa_\varepsilon,\kappa_\mu\right) \;=\; \sum_{q=0}^{n-1}\kappa_\mu^q(1-\kappa_\mu)^{n-1-q}\binom{n-1}{q}\sum_{r\in R^f}A^f\cdot\Pi^{f1}\left[M^f(r,q,m),\kappa_\varepsilon\right], \tag{A22}$$

$$\mathcal{W}^c\left(m,\kappa_\varepsilon,\kappa_\mu\right) \;=\; \sum_{q=0}^{n-1}\kappa_\mu^q(1-\kappa_\mu)^{n-1-q}\binom{n-1}{q}\sum_{r\in Rc}A^c\cdot\Pi^{c1}\left[M^c(r,q,m),\kappa_\varepsilon\right]. \tag{A23}$$

Lemma 1 *It holds*

(i) $\sum_{r\in R^f}A^f\cdot M^f(r,q,m) = \underbrace{\left(m-\frac{m}{n-1}q\right)}_{=:B^f(q)}$ *and* $\sum_{r\in R^c}A^c\cdot M^c(r,q,m) = \underbrace{\left(m+\frac{n-m}{n-1}q\right)}_{:=B^c(q)}$.

(ii) $\sum_{r\in R^f}A^f\cdot\left[M^f(r,q,m)\right]^2 = \underbrace{\left[m^2+\frac{m(n+3m-2mn-1)q+m(m-1)q^2}{(n-1)(n-2)}\right]}_{=:G^f(q,q^2)}$ *and*

$\sum_{r\in R^c}A^c\cdot\left[M^c(r,q,m)\right]^2 = \underbrace{\left[m^2+\frac{(n-m)(2mn-3m-1)q+(n-m)(n-m-1)q^2}{(n-1)(n-2)}\right]}_{=:G^c(q,q^2)}$.

Proof of (i): Observe that

$$\sum_{r\in R^f}A^f\cdot M^f(r,q,m) = \sum_{r\in R^f}A^f\cdot\left[(1+q)\underbrace{s_i^f}_{=0}+(n-m-1-r)\underbrace{s_a^f}_{=0}+(m-q+r)\underbrace{s_a^c}_{=1}\right] = \sum_{r\in R^f}A^f\cdot(m-q+r), \tag{A24}$$

$$\sum_{r\in R^c}A^c\cdot M^c(r,q,m) = \sum_{r\in R^c}A^c\cdot\left[(1+q)\underbrace{s_i^c}_{=1}+(m-1-r)\underbrace{s_a^c}_{=1}+(n-m-q+r)\underbrace{s_a^f}_{=0}\right] = \sum_{r\in R^c}A^c\cdot(m+q-r). \tag{A25}$$

Applying Identity 2.1 of Mestrovic (2018) it holds

$$\sum_{r\in R^f}A^f\cdot(q-r) = \frac{m}{n-1}q, \quad \sum_{r\in R^c}A^c\cdot(q-r) = \frac{n-m}{n-1}q \tag{A26}$$

which implies

$$\sum_{r\in R^f} A^f \cdot (m-q+r) = \left(m - \frac{m}{n-1}q\right), \quad \sum_{r\in R^c} A^c \cdot (m+q-r) = \left(m + \frac{n-m}{n-1}q\right). \tag{A27}$$

Proof of (ii): Observe that

$$\sum_{r\in R^f} A^f \cdot \left[M^f(r,q,m)\right]^2 = \sum_{r\in R^f} A^f \cdot [(m-q+r)]^2, \quad \sum_{r\in R^c} A^c \cdot [M^c(r,q,m)]^2 = \sum_{r\in R^c} A^c \cdot [(m+q-r)]^2. \tag{A28}$$

Applying Identity 2.17 of Mestrovic (2018) it holds

$$\sum_{r\in R^f} A^f \cdot (q-r)^2 \;=\; \frac{mq}{(n-1)(n-2)}\;[(n-m-1)+(m-1)q], \tag{A29}$$

$$\sum_{r\in R^c} A^c \cdot (q-r)^2 \;=\; \frac{(n-m)q}{(n-1)(n-2)}\;[m-1+(n-m-1)q]. \tag{A30}$$

Next, we calculate

$$\begin{aligned}\sum_{r\in R^f} A^f \cdot [(m-q+r)]^2 &= \sum_{\substack{r\le n-m-1\\ q-r\le m}} A^f \cdot [m-(q-r)]^2\\ = \sum_{r\in R^f} A^f \cdot \left[m^2 - 2m(q-r) + (q-r)^2\right] &= \left\{m^2 + \frac{m(n+3m-2mn-1)q + m(m-1)q^2}{(n-1)(n-2)}\right\}\end{aligned} \tag{A31}$$

and

$$\begin{aligned}\sum_{r\in R^c} A^c \cdot [(m+q-r)]^2 &= \sum_{\substack{r\le m-1\\ q-r\le n-m}} A^c \cdot [m+(q-r)]^2\\ = \sum_{r\in R^c} A^c \cdot \left[m^2 + 2m(q-r) + (q-r)^2\right] &= \left\{m^2 + \frac{(n-m)(2mn-3m-1)q + (n-m)(n-m-1)q^2}{(n-1)(n-2)}\right\}.\end{aligned} \tag{A32}$$

■

Taking advantage of Lemma 1 in (A22) and (A23) we get

$$\begin{aligned}
\mathcal{W}^f\left(m,\kappa_\varepsilon,\kappa_\mu\right) &= \sum_{q=0}^{n-1}\kappa_\mu^q(1-\kappa_\mu)^{n-1-q}\binom{n-1}{q}\Bigg\{\frac{\alpha^2-2\alpha n\delta+\delta^2\big[(2n-1)(1-\kappa_\varepsilon)^2+2\left(G^f(q,q^2)-B^f(q)\right)(1-\kappa_\varepsilon)\big]}{2\beta}\,, \\
&= \qquad\qquad\qquad\qquad\qquad + \frac{\alpha^2-2\alpha n\delta+\delta^2\big[+n^2(2-\kappa_\varepsilon)\kappa_\varepsilon\big]}{2\beta}\Bigg\}, \qquad (A33) \\
\mathcal{W}^c\left(m,\kappa_\varepsilon,\kappa_\mu\right) &= \sum_{q=0}^{n-1}\kappa_\mu^q(1-\kappa_\mu)^{n-1-q}\binom{n-1}{q}\Bigg\{\frac{\alpha^2-2\alpha n\delta-\delta^2\big[2B^c(q)(1-\kappa_\varepsilon)(1+n\kappa_\varepsilon)-G^c(q,q^2)(1-\kappa_\varepsilon^2)}{2\beta} \\
&\qquad\qquad\qquad\qquad\qquad + \frac{\alpha^2-2\alpha n\delta-\delta^2\big[n(n\kappa_\varepsilon^2-2(n-1)\kappa_\varepsilon-2)\big]}{2\beta}\Bigg\}. \qquad (A34)
\end{aligned}$$

Finally, we obtain

$$\begin{aligned}
\mathcal{W}^f\left(m,\kappa_\varepsilon,\kappa_\mu\right) &= \frac{\alpha^2-2\alpha n\delta+\delta^2\big[(2n-1)(1-\kappa_\varepsilon)^2+2\left(D^f(\mathbf{E}[X],\mathbf{E}[X]^2)-B^f(\mathbf{E}[X])\right)(1-\kappa_\varepsilon)+n^2(2-\kappa_\varepsilon)\kappa_\varepsilon\big]}{2\beta}, \qquad (A35) \\
\mathcal{W}^c\left(m,\kappa_\varepsilon,\kappa_\mu\right) &= \frac{\alpha^2-2\alpha n\delta-\delta^2\big[2B^c(\mathbf{E}[X])(1-\kappa_\varepsilon)(1+n\kappa_\varepsilon)-D^c(\mathbf{E}[X],\mathbf{E}[X]^2)(1-\kappa_\varepsilon^2)+n(n\kappa_\varepsilon^2-2(n-1)\kappa_\varepsilon-2)\big]}{2\beta}, \qquad (A36)
\end{aligned}$$

where X is a random variable that follows the binomial distribution with parameter $n-1$ and probability $p = \kappa_\mu$. Observe that $\mathbf{E}[X] = (n-1)\kappa_\mu$ and $\mathbf{E}[X^2] = (n-1)\kappa_\mu(1-\kappa_\mu) + (n-1)^2\kappa_\mu^2$.

C: The derivatives of the functions $\Pi^{h1}(m, \kappa_\varepsilon)$:

Differentiation of $\Pi^{h1}(m, \kappa_\varepsilon)$ from (15) with respect to m yields

$$\Pi_m^{f1} = -D' \cdot \left[\mathcal{E}^c(m,\kappa_\varepsilon) - \mathcal{E}^f(m,\kappa_\varepsilon) + m\mathcal{E}_m^c\right] > 0, \tag{A37}$$

$$\Pi_m^{c1} = (B' - mD')\,\mathcal{E}_m^c - D' \cdot \left[\mathcal{E}^c(m,\kappa_\varepsilon) - \mathcal{E}^f(m,\kappa_\varepsilon)\right]. \tag{A38}$$

Inserting $B'(e_i^c) - \delta m = \delta(n-m)\kappa_\varepsilon$ from (A13), $\mathcal{E}_m^c = -\frac{(1-\kappa_\varepsilon)\delta}{\beta}$ and (14) into (A38) we obtain

$$\begin{aligned}\Pi_m^{c1} &= -\frac{\delta(n-m)\kappa_\varepsilon(1-\kappa_\varepsilon)\delta^2}{\beta} + \delta\left[\frac{(1-\kappa_\varepsilon)(m-1)\delta}{\beta}\right] \\ &= -\left[(n-m)\kappa_\varepsilon - (m-1)\right]\frac{(1-\kappa_\varepsilon)\delta^2}{\beta}.\end{aligned} \tag{A39}$$

From (A39) we infer

$$\Pi_m^{c1} \lesseqqgtr 0 \quad \Longleftrightarrow \quad \kappa_\varepsilon \gtreqqless \frac{m-1}{n-m}. \tag{A40}$$

D: Derivation of (24) and Proof of Proposition 1:

For $\kappa_\mu = 0$, the stability function turns into

$$S(m,\kappa_\varepsilon,0) = \frac{(m-1)(1-\kappa_\varepsilon)\delta^2\;\left[3 - m(1-\kappa_\varepsilon) - (2n-1)\kappa_\varepsilon\right]}{2\beta}. \tag{A41}$$

Observe that $S_m(m,\kappa_\varepsilon,0) = -\frac{(1-\kappa_\varepsilon)^2\delta^2}{2\beta} < 0$. Suppose $\kappa_\varepsilon < \frac{1}{2n-3}$. Solving $S(m,\kappa_\varepsilon,0) = 0$ with respect to m yields

$$m = \mathcal{M}(\kappa_\varepsilon, 0) = \frac{3-(2n-1)\kappa_\varepsilon}{1-\kappa_\varepsilon} \quad \text{if}\,\kappa_\varepsilon < \frac{1}{2n-3}. \tag{A42}$$

Suppose $\kappa_\varepsilon \geq \frac{1}{2n-3}$. Due to $S_m(m,\kappa_\varepsilon,0) < 0$ for $m \in]1,n]$ it holds

$$m = \mathcal{M}(\kappa_\varepsilon, 0) = 1 \quad \text{if}\,\kappa_\varepsilon \geq \frac{1}{2n-3}. \tag{A43}$$

Proposition 1(i) follows from setting $\kappa_\varepsilon = 0$ in (A42) and Proposition 1(ii) follows from $\mathcal{M}_{\kappa_\varepsilon} = -\frac{2(n-2)}{(1-\kappa_\varepsilon)^2} < 0$ and $m = 2 \quad \Longleftrightarrow \quad \kappa_\varepsilon = \frac{1}{2n-3}$.

E: Proof of Proposition 2:

For $\kappa_\varepsilon = 0$, the stability function turns into

$$S\left(m, 0, \kappa_\mu\right) = \frac{\delta^2 F^\mu(m, \kappa_\mu)}{2\beta}, \tag{A44}$$

where

$$F^\mu(m, \kappa_\mu) \quad = \quad (n^2 - n - 4)\kappa_\mu^2 - (n-8)\kappa_\mu - m^2(1-\kappa_\mu)^2 - 3 + m(1-\kappa_\mu)[(2n-7)\kappa_\mu - 4]. \tag{A45}$$

The function F^μ satisfies $F^\mu(1, \kappa_\mu) = (n-1)^2 > 0$ and $F^\mu_{mm}(m, \kappa_\mu) = -2(1-\kappa_\mu)^2 < 0$. Solving $F^\mu(m, \kappa_\mu) = 0$ with respect to m yields

$$m = \frac{-4 + \kappa_\mu^2(2n-7) - \kappa_\mu(2n-11) + (1-\kappa_\mu)\sqrt{4 + \kappa_\mu(12n-24) + \kappa_\mu^2(8n^2 - 32n + 33)}}{2(1-\kappa_\mu)^2}. \tag{A46}$$

(ii) Solving $F^\mu(n, \kappa_\mu) = 0$ with respect to κ_μ we get

$$\kappa_\mu = 1 - \frac{\sqrt{2}\sqrt{(n-2)(n-1)}}{2(n-2)} =: \tilde{\kappa}_\mu. \tag{A47}$$

Due to $F^\mu(n, 0) = -(n-3)(n-1) < 0$ and $F^\mu_{\kappa_\mu \kappa_\mu}(n, \kappa_\mu) = -4(n-2)(n-1) < 0$ and $F^\mu(n, 1) = (n-1)^2 > 0$, the grand coalition is stable if $\kappa_\mu > \tilde{\kappa}_\mu$. Figure 8 illustrates $\tilde{\kappa}_\mu$ in dependence of n. Observe that $\lim_{n\to\infty} \tilde{\kappa}_\mu = 1 - \frac{1}{\sqrt{2}}$

(i) Summarizing our results, the stable coalition is determined by

$$\mathcal{M}(0, \kappa_\mu, n) = \begin{cases} \frac{-4+\kappa_\mu^2(2n-7)-\kappa_\mu(2n-11)+(1-\kappa_\mu)\sqrt{4+\kappa_\mu(12n-24)+\kappa_\mu^2(8n^2-32n+33)}}{2(1-\kappa_\mu)^2} & \text{if } 0 \leq \kappa_\mu < \tilde{\kappa}_\mu, \\ n & \text{if } \tilde{\kappa}_\mu \leq \kappa_\mu < 1. \end{cases} \tag{A48}$$

The poof that $\mathcal{M}(0, \kappa_\mu, n)$ increases from $\mathcal{M}(0, 0, n) = 3$ to $\mathcal{M}(0, \tilde{\kappa}_\mu, n) = n$ is provided for $n = 200$ in Figure 9.

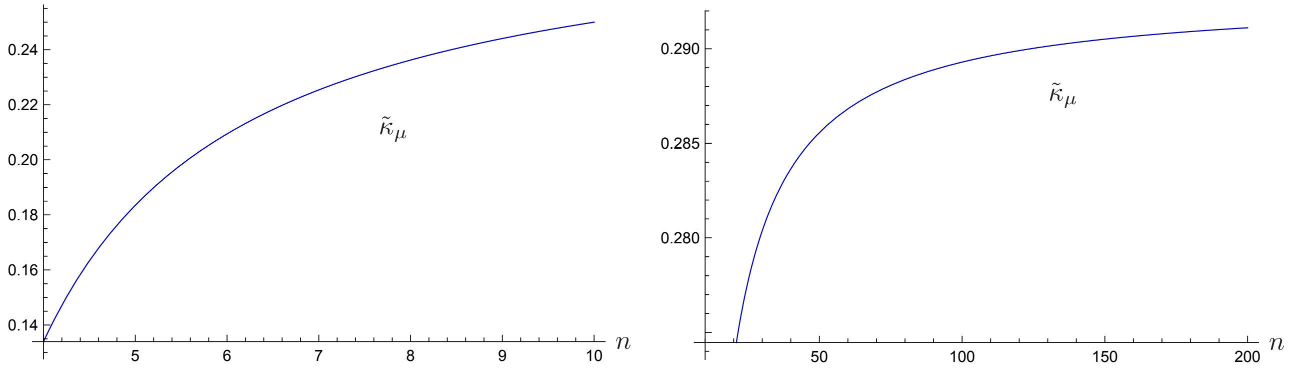

Figure 8: $\tilde{\kappa}_\mu$ in dependence of n

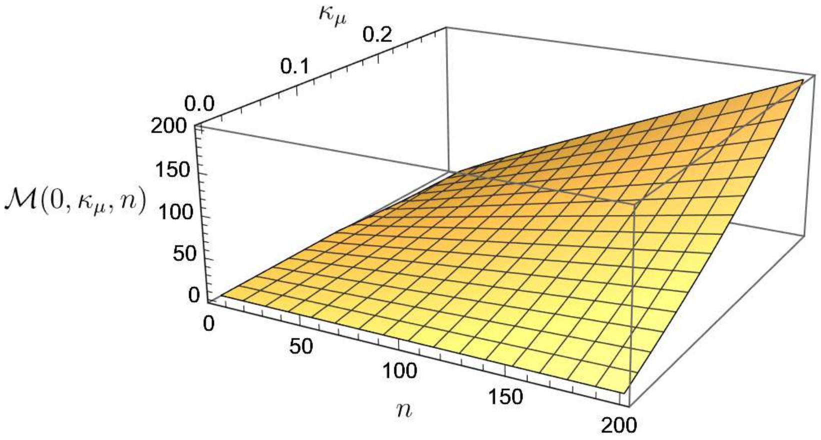

Figure 9: The function $\mathcal{M}(0,\kappa_\mu,n)$

F: Proof of Proposition 3 (iii):

For $\kappa_\varepsilon = \kappa_\mu =: \kappa$, the stability function is given by

$$S\left(m,\kappa,\kappa\right) = \frac{\delta^2 F^\kappa(m,\kappa)}{2\beta}, \tag{A49}$$

where

$$F^\kappa(m,\kappa) \quad = \quad -(1-\kappa)^2\left[3+m^2(1-\kappa)^2-(n+4)\kappa+n(n-1)\kappa^2+m(-4+7\kappa-(2n-1)\kappa^2)\right]. \tag{A50}$$

(Due to $F^\kappa(0,n) = -(n-3)(n-1) < 0$, $F^\kappa_\kappa(\kappa,n) = 2(n-2)(n-1) > 0$ and

$$F^\kappa(\tilde{\kappa},n) = 0 \qquad \Longleftrightarrow \tilde{\kappa} = \frac{n-3}{2(n-2)}. \tag{A51}$$

the grand coalition is stable if and only if $\kappa \geq \tilde{\kappa}$.

$F^\kappa(m,\kappa) = 0$ is a quadratic equation in m, its discriminant is given by

$$D(\kappa,n) := 4+4(n-4)\kappa-(3-12n+4n^2)\kappa^2+(30-32n+8n^2)\kappa^3+\kappa^4 \tag{A52}$$

and is illustrated in Figure 10. The blue plane in Figure 10 is the plane $(0,\kappa,n)$. The thresholds $\underline{\kappa}$ and $\bar{\kappa}$ are implicitly determined by $D(\underline{\kappa},n) = 0$, $D(\bar{\kappa},n) = 0$ and are shown in Figure 11.

Suppose that $n \in]10,200]$. For $\underline{\kappa} < \kappa < \underset{\sim}{\kappa}$ the discriminant is negative such that the quadratic equation $F^\kappa(m,\kappa) = 0$ has no solution and no stable coalition exists. For $\kappa < \underline{\kappa}$ or $\underset{\sim}{\kappa} < \kappa < \tilde{\kappa}$, the discriminant is positive and has two solutions. The relevant solution is given by

$$m = \frac{2-4\kappa-n^2\kappa^2-n(2-3\kappa-2\kappa^2)+\sqrt{(n-1)^2[1+(n-3)\kappa-(n^2-3n+2)\kappa^2+2(n-2)^2\kappa^3]}}{(1-\kappa)[1-2\kappa-n(1-\kappa)]}. \tag{A53}$$

Summarizing our results, the stable coalition is determined by the function

$$\mathcal{M}(\kappa,\kappa,n) = \begin{cases} \frac{2-4\kappa-n^2\kappa^2-n(2-3\kappa-2\kappa^2)+\sqrt{(n-1)^2[1+(n-3)\kappa-(n^2-3n+2)\kappa^2+2(n-2)^2\kappa^3]}}{(1-\kappa)[1-2\kappa-n(1-\kappa)]} & \text{if } 0 \leq \kappa < \underline{\kappa}, \\ 1 & \text{if } \underline{\kappa} < \kappa < \underset{\sim}{\kappa}, \\ \frac{2-4\kappa-n^2\kappa^2-n(2-3\kappa-2\kappa^2)+\sqrt{(n-1)^2[1+(n-3)\kappa-(n^2-3n+2)\kappa^2+2(n-2)^2\kappa^3]}}{(1-\kappa)[1-2\kappa-n(1-\kappa)]} & \text{if } \underset{\sim}{\kappa} \leq \kappa < \tilde{\kappa} \equiv \frac{n-3}{2(n-2)}, \\ n & \text{if } \tilde{\kappa} \leq \kappa < 1. \end{cases} \tag{A54}$$

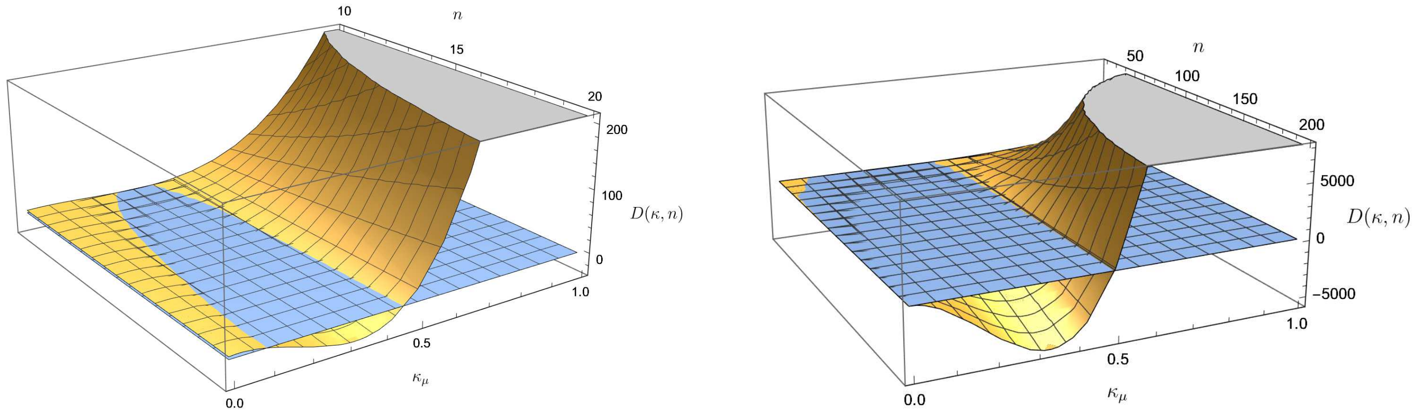

Figure 10: The discriminant for $n \in]9, 20]$ (left panel) and $n \in [20, 200]$ (right panel)

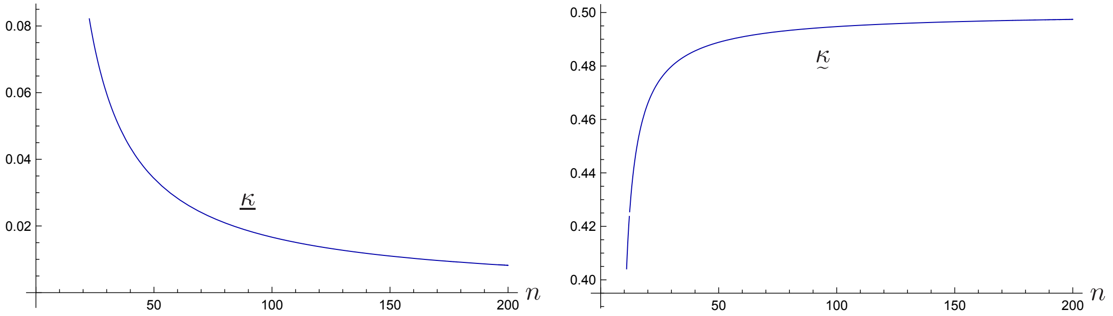

Figure 11: The bounds $\underline{\kappa}$ and $\underset{\sim}{\kappa}$ in dependence of n

The function $\mathcal{M}(\kappa, \kappa, n)$ is plotted for $0 \le \kappa < \underline{\kappa}$ and $\underset{\sim}{\kappa} \le \kappa < \tilde{\kappa}$ in Figure 12 and 13, respectively.

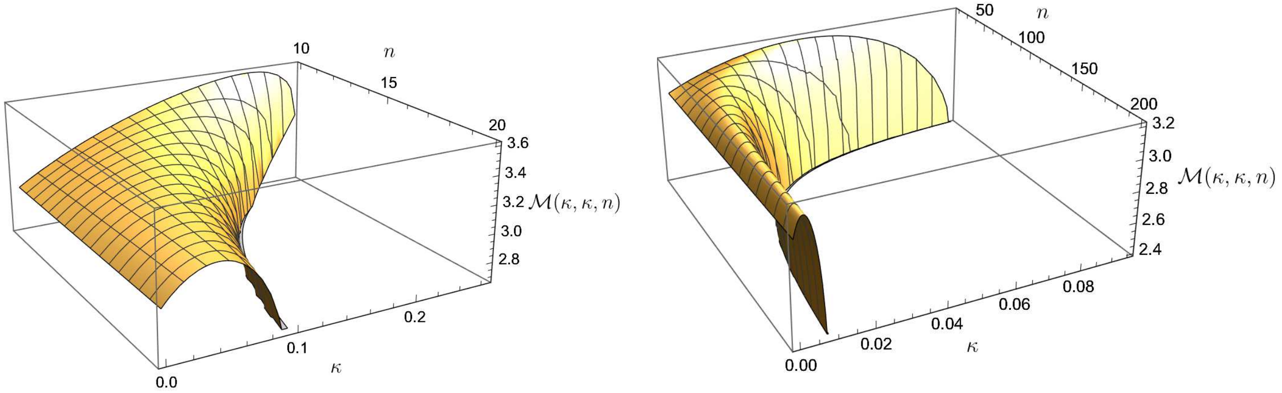

Figure 12: $\mathcal{M}(\kappa, \kappa, n)$ for $0 \leq \kappa < \underline{\kappa}$ and $n \in]9, 200]$

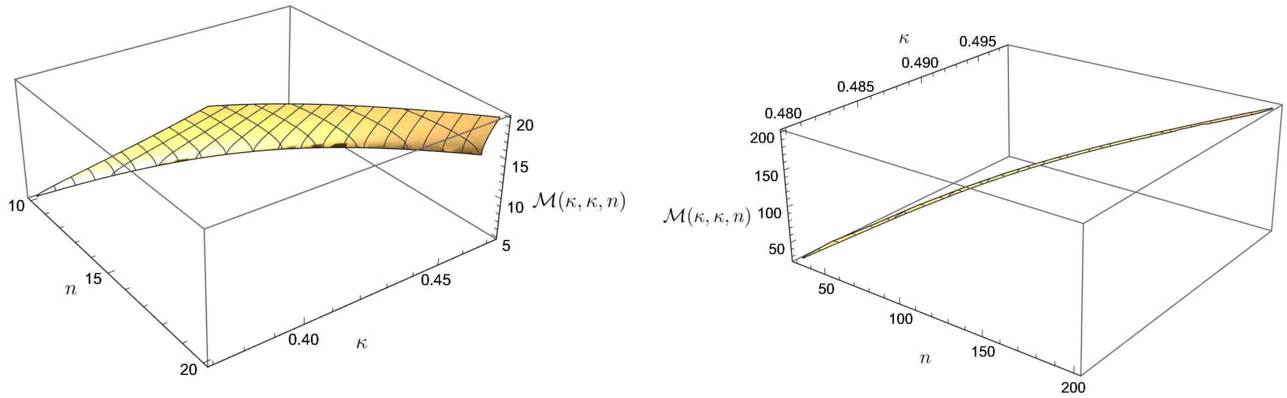

Figure 13: $\mathcal{M}(\kappa, \kappa, n)$ for $\underset{\sim}{\kappa} \leq \kappa < \tilde{\kappa}$ and $n \in]9, 200]$

Finally, suppose that $n \in [3, 9]$. Then the discriminant is positive for all $\kappa \in \left[0, \frac{n-3}{2(n-2)}\right]$ and the stable coalition is characterized by

$$\mathcal{M}(\kappa, \kappa, n) = \begin{cases} \frac{2-4\kappa-n^2\kappa^2-n(2-3\kappa-2\kappa^2)+\sqrt{(n-1)^2[1+(n-3)\kappa-(n^2-3n+2)\kappa^2+2(n-2)^2\kappa^3]}}{(1-\kappa)[1-2\kappa-n(1-\kappa)]} & \text{if } 0 \leq \kappa < \tilde{\kappa} \equiv \frac{n-3}{2(n-2)}, \\ n & \text{if } \tilde{\kappa} \leq \kappa < 1. \end{cases} \tag{A55}$$

$\mathcal{M}(\kappa, \kappa, n)$ is illustrated in Figure 14.

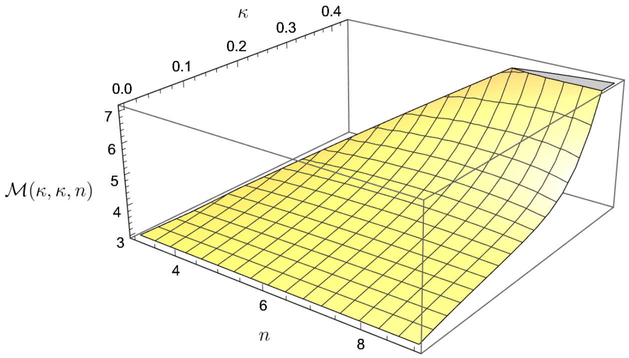

Figure 14: $\mathcal{M}(\kappa, \kappa, n)$ for $\kappa \in \left[0, \frac{n-3}{2(n-2)}\right]$ and $n \in [3, 9]$

References

Aghion, P., Benabou, R., Martin, R. and A. Roulet (2023): Environmental preferences and technological choices: Is market competition clean or dirty?, *American Economic Review: Insights* 5, 1-20.

Alger, J. and and J-F. Laslier (2022): Homo moralis goes to the voting booth: coordination and information aggregation, *Journal of Theoretical Politics* 34, 280-312.

Alger, J. and J.W. Weibull (2020): Morality: evolutionary foundations and policy implications, in Basu, K., Rosenblatt, D. and C. Sepulveda (eds.), *The State of Economics, the State of the World*, MIT Press.

Alger, J. and J.W. Weibull (2017): Strategic behavior of altruists and moralists, *Games* 8, 1-21.

Alger, J. and J.W. Weibull (2016): Evolution and Kantian morality, *Games and Economic Behavior* 98, 56-67.

Alger, J. and J.W. Weibull (2013): Homo moralis - preference evolution under incomplete information and assortative matching, *Econometrica* 81, 2269-2302.

Ambec, S. and P. De Donder (2022): Environmental policy with green consumerism, *Journal of Environmental Economics and Management* 111, 102584.

Bolton, G. E and A. Ockenfels (2000): ERC: A theory of equity, reciprocity, and competition, *American Economic Review* 91, 166-193.

Barrett, S. (1994): Self-enforcing international environmental agreements, *Oxford Economic Papers* 46, 878-894.

Buchholz, W., Peters, W. and A. Ufert (2018): International environmental agreements on climate protection: A binary choice model with heterogeneous agents, *Journal of Economic Behavior and Organization* 154, 191-205.

Carraro, C. and D. Siniscalco (1993): Strategies for the international protection of the environment, *Journal of Public Economics* 52, 309-328.

Carraro, C. and D. Siniscalco (1991): Strategies for the international protection of the environment, CEPR Discussion Paper 568.

Daube, M. and D. Ulph (2016): Moral behaviour, altruism and environmental policy, *Environmental and Resource Economics* 63, 505-522.

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

Eichner, T. and R. Pethig (2024): International environmental agreements when countries behave morally, *Journal of Environmental Economics and Management* 125, 102955.

Eichner, T. and R. Pethig (2021): Climate policy and moral consumers, *Scandinavian Journal of Economics* 123, 1190-1226.

Ellen, V. D. W., Steg, L. and K. Keizer, K. (2013): It is a moral issue: The relationship between environmental self-identity, obligation-based intrinsic motivation and pro-environmental behavior, *Global Environmental Change* 23, 1258-1265.

Fehr, E. and K. M. Schmidt (1999): A theory of fairness, competition, and cooperation, *Quarterly Journal of Economics* 114, 817-868.

Finus, M. and S. Maus (2008): Modesty may pay!, *Journal of Public Economic Theory* 10, 801-826.

Grafton, R.Q., Kompas, T. and N. van Long (2017): A brave new world? Kantian-Nashian interaction and the dynamics of global climate change mitigation, *European Economic Review* 99, 31-42.

Herweg, F. and K. M. Schmidt (2022): How to regulate carbon emissions with climate-conscious consumers, *Economic Journal* 132, 2992-3019.

Hoel, M. (1992): International environmental conventions: The case of uniform reductions of emissions, *Environmental and Resource Economics* 2, 141-159.

Kant, I. (1785): *Grundlegung zur Metaphysik der Sitten.* Trad: Mary, G. and J. Timmermann (2011) *Groundwork of the Metaphysics of Morals: A German-English Edition.* Cambridge, Mass.: Cambridge University Press.

Laffont, J.-J. (1975): Macroeconomic constraints, economic efficiency and ethics: An introduction to Kantian economics, *Economica* 42, 430-437.

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

Liobikiene, G., Mandravickaite, J. and J. Bernatoniene (2016): Theory of planned behavior approach to understand the green purchasing behavior in the EU: A cross-cultural study, *Ecological Economics* 125, 38-46.

Mestrovic, R. (2018): Several generalizations and variations of Chu-Vandermonde identity, arXiv preprint arXiv:1807.10604.

Nyborg, K. (2018): Reciprocal climate negotiators, *Journal of Environmental Economics and Management* 92, 707-725.

Roemer, J.E. (2015): Kantian optimization. A microfoundation for cooperation, *Journal of Public Economics* 127, 45-57.

Roemer, J.E. (2010): Kantian equilibrium, *Scandinavian Journal of Economics* 112, 1-24.

Rubio, S.J. and A. Ulph (2006): Self-enforcing agreements and international trade in greenhouse emission rights, *Oxford Economic Papers* 58, 233-263.

Schopf, M. (2023) : Self-enforcing International Environmental Agreements and Altruistic Preferences, Beiträge zur Jahrestagung des Vereins für Socialpolitik 2023: Growth and the "sociale Frage", ZBW - Leibniz Information Centre for Economics, Kiel, Hamburg.

Ulph, A. and D. Ulph (2023): International cooperation and Kantian moral behaviour - complements or substitutes?, Economics Discussion Paper Series EDP-2302, The University of Manchester.

Van Long, N. (2020): A dynamic game with interaction between Kantian players and Nashian players, in Pineau, P.-O., Sigué, S. and S. Taboubi (eds.), *Games in Management Science: Essays in Honor of Georges Zaccour*, Springer: Switzerland.

Van der Pol, T., Weikard, H.-P. and E. van Ierland (2012): Can altruism stabilize international climate agreements, *Ecological Economics* 81, 33-59.

Vogt, C. (2016): Climate coalition formation when players are heterogeneous and inequality averse, *Environmental and Resource Economics* 65, 33-39.