CESifo WORKING PAPERS

10090
2022

November 2022

International Environmental Agreements When Countries Behave Morally

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. 10090*

International Environmental Agreements When Countries Behave Morally

Abstract

In the standard theoretical literature on forming international environmental agreements (IEAs) countries use to be self-interested materialists and stable coalitions are small. This paper analyzes IEA games with countries that exhibit Kantian moral behavior. Countries may behave morally with respect to both emissions (reduction) and membership in an IEA. If countries are emissions Kantians or membership Kantians the outcome of the corresponding IEA games is socially optimal. To model more realistic Kantian behavior, we define an emissions [membership] moralist as a country whose welfare is the weighted average of the welfare of an emissions [membership] Kantian and a materialist. The game with emissions moralists produces stable coalitions not larger than those in the standard game with materialists. The game with membership moralists yields stable coalitions that are increasing in the membership morality. Finally, we consider countries who are moderate moralists with respect to both emissions and membership. In that encompassing IEA game the size of the coalition is increasing in the emissions morality, the membership morality, and in the weight of the membership moralist's welfare. Depending on parameter values, the grand coalition may or may not be attained if one of the moral parameter increases and tends towards one.

JEL-Codes: C720, Q500, Q580.

Keywords: international environmental agreement, stable coalitions, moral behavior, Kantian ethics.

Thomas Eichner
Department of Economics
University of Hagen
Universitätsstr. 41
Germany – 58097 Hagen
thomas.eichner@fernuni-hagen.de

Rüdiger Pethig
Department of Economics
University of Siegen
Unteres Schloss 3
Germany – 57072 Siegen
pethig@vwl.wiwi.uni-siegen.de

Financial support from the German Science Foundation (DFG grant number EI 847/2-1) is gratefully acknowledged. We thank Alistair Ulph and David Ulph for numerous intensive and stimulating discussions on all aspects of our paper and gratefully acknowledge their helpful comments. Remaining errors are the authors' sole responsibility.

1 Introduction

Climate scientists predict severe future climate damage with high probability, unless global emissions of greenhouse gases per annum are reduced down to net zero emissions by 2050 (UNEP 2019). In the Paris climate agreement, the international community responds to that challenge by agreeing on an ambitious goal to cut global emissions through nationally determined emissions reductions. Given the current large 'emissions gap', meeting the Paris goal requires all governments to strongly intensify their mitigation efforts. Governments, in turn, will pursue more stringent climate policies to the extent only that their electorate induces them to increase their mitigation effort and is willing to bear the pertaining mitigation costs (Bernauer et al. 2016). Ellen et al. (2013) and Liobikiene et al. (2016) provide evidence that increasing numbers of individuals deliberately reduce their carbon footprint below the level self-interested consumers would choose. These individuals want their government (i) to curb domestic emissions more effectively and (ii) to play a more pro-active part in the ongoing Paris process. Here we suppose governments respond to their pro-climate consumer-voters by stepping up mitigation and international cooperation.

Explanations of the consumers' motivation to plead for more stringent climate policies and international cooperation have been suggested along two different lines. One line is to amend self-interested preferences by including arguments such as altruism, equality, fairness or warm glow.[1] The other line, which we will pursue here, is an approach with a flavor of Kantian ethics (Kant 1785) the core of which is the categorical imperative. It says that one should take (or recommend) 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 individuals. The present paper considers countries exhibiting Kantian moral behavior with respect to emissions and membership in an international environmental agreement (IEA) and aims to investigate the impact of that behavior on the formation of an IEA.

We formalize Kantian moral behavior along the lines of Alger and Weibull (2013, 2016, 2020) with the qualification that they focus on moral individuals while we consider moral governments/countries engaged in forming an IEA. Our benchmark is the canonical game of

[1]See Van Long (2016), Dasgupta et al. (2016), Nyborg (2018a), and the literature cited therein. In an experimental study, Jakob et al. (2017) find that participants feel morally responsible for cleaning up (environmental) externalities they caused, even if delegating the task would be more efficient. Their experiment disregards the conflict between individual and collective rationality, which is at the core of (global) public good analyses (and experiments) such as climate change mitigation.

the early IEA literature with conventional self-interested or *materialistic* countries (Barrett 1994, Carraro and Siniscalco 1993, Hoel 1992) that is known to have stable coalitions with no more than three countries (under restrictive conditions necessary to characterize its solution). Here we will present a number of IEA games with different specifications of moral behavior listed in Table 1. Our principal interest is in the last game in Table 1, Game $\varepsilon\mu\sigma$, where the players are countries we call *general moralists*. The behavior of these countries is quite complex, because it consists of various facets of moral and materialist behavior. All other games in Table 1 serve to understand the components of the general moralist's behavior, and we will discuss them, one at a time.

Self-interested or **materialistic countries**	(Section 2.1)	**Game** σ
Kantians w.r.t emissions	(Section 2.2):	
Rigorous Kantians w.r.t. emissions	= **emissions Kantians**	**Game** ε
Blend of emissions Kantians		
and materialistic countries	= **emissions moralists**	**Game** $\varepsilon\sigma$
Kantians w.r.t. membership	(Section 2.3):	
Rigorous Kantians w.r.t. membership	= **membership Kantians**	**Game** μ
Blend of emissions Kantians		
and materialistic countries	= **membership moralists**	**Game** $\mu\sigma$
Moralists w.r.t. emissions *and* membership	(Section 2.4):	
	= **general moralists**	**Game** $\varepsilon\mu\sigma$

Table 1: IEA games with different specifications of moral countries

The key principles of Kantian morality are expressed in the behavior of *emissions Kantians* (in Game ε in Table 1) and *membership Kantians* (in Game μ). An emissions Kantian considers its (moral) welfare to be the welfare a *materialist* enjoys on the counterfactual assumption that all other countries choose the same emissions. Accordingly, an emissions Kantian's welfare is very low when its own emissions are very high on the assumption that all other countries would choose theses high emissions as well. It is then also clear that the emissions Kantian's welfare increases rapidly, when it curbs the emissions assuming that all others follow up. Its choice of emissions is then in the spirit of the Kantian categorical imperative which requires to choose that level of emissions it would advocate all other countries choose as well. A membership Kantian who is [not] a member of a given coalition considers its (moral) welfare to be the welfare a materialist enjoys when it is [not] in a coalition on the

counterfactual assumption that all other countries are [not] in the coalition. So, as a member of a (small or large) given coalition the membership Kantian acts as if it is a member of the grand coalition and if it is not a member of the given coalition it acts as if there is no coalition at all. Its (moral) welfare in the former state is obviously higher than in the latter such that the categorical imperative leads it to advocate all countries to join the coalition.

It is straightforward to show that the solution to the Games ε and μ is the social optimum that is reached in the grand coalition. In Game ε, it is reached even without the need of forming a coalition. However, the morality of emissions and membership Kantians is so rigorous that such individuals, let alone countries, may hardly be found in the real world. Following Alger and Weibull (2013, 2016, 2020) we therefore introduce less rigorous Kantian behavior by defining an *emissions moralist* in Game $\varepsilon\sigma$ and a *membership moralist* in Game $\mu\sigma$. The former [latter] considers the moral value of domestic emissions [of its choice of membership] to be the convex combination[2] of a materialist's and an emissions Kantian's[3] [a membership Kantian's] welfare. The weight between zero and one attached to the emissions [membership] Kantian's welfare is denoted degree of emissions [membership] morality.[4] Unexpectedly, in Game $\varepsilon\sigma$ the climate coalition turns out to consist of three countries only, like in Game σ in Table 1, where countries are materialists. So, emissions moralists of the Alger-Weibull type do not improve the coalition formation process. In Game $\mu\sigma$ the size of the coalition is increasing in the degree of membership morality, which conforms to intuition.

The moderate Kantian moral behavior in the Games $\varepsilon\sigma$ and $\mu\sigma$ is clearly more realistic than the rigorous Kantian moralism in the Games ε and μ. However, it would be onesided and therefore unsatisfactory to consider moralists to be either emissions moralist only or membership moralists only. To account for moral behavior 'on both dimensions', we finally suggest the Game $\varepsilon\mu\sigma$ with countries whose welfares/payoffs are a convex combination of the welfare of an emissions moralist and a membership moralist.[5] Our main result is that in the encompassing Game $\varepsilon\mu\sigma$ the size of the coalition is increasing in the degree of emissions

[2]The convex combination of x and y is $z := \lambda x + (1-\lambda)y$ where $\lambda \in [0,1]$ is the weight on x.

[3]Eichner and Pethig (2021) apply the morality concept of Alger and Weibull for emissions in the context of international emissions cap competition, and, departing from Alger and Weibull and the present paper, they assume that agents differ with respect to their degree of emissions morality.

[4]If the degrees of emissions or membership morality take on an extreme value zero or one, the Game $\varepsilon\sigma$ or $\mu\sigma$ degenerate into one of the Games ε, μ or σ.

[5]The perfectly symmetric moralist is included as that special case in which the weight of the emissions moralist's and the membership moralist's welfare is 0.5 and where the degrees of emissions morality and membership morality are the same.

morality, the degree of membership morality, and also in the moral parameter that increases the weight of the membership moralist's welfare. The grand coalition, and with it the social optimum, is attained when the values of the moral parameters are sufficiently large but they may still be significantly smaller than one.

The present paper contributes to two different strands of the literature. The first is the literature on self-enforcing IEAs. In the basic model of the early IEA literature alluded to above that has been further studied by Rubio and Ulph (2006) and Diamantoudi and Sartzetakis (2006), the stable coalition consists of at most three countries no matter whether the coalition plays Nash or Stackelberg. The robustness of this pessimistic result has been examined in various different extensions of the canonical model. Among these extensions are modesty on the part of the coalition (Finus and Maus 2008), adaptation and mitigation (Bayramoglu et al. 2018), ancilliary effects (Finus and Rübbelke 2013), inequality aversion (Lange and Vogt 2003, Vogt 2016), altruism (van der Pol et al. 2012) and reciprocity (Nyborg 2018b and Buchholz et al. 2018). Furthermore, Eichner and Pethig (2013, 2015) have added trade, McEvoy and McGinty (2018) have studied emissions taxes, Ansik et al. (2019) have investigated support from outsiders and De Zeeuw (2018) and Diamantoudi and Sartzetakis (2016) have investigated foresighted countries. While inequality aversion, ancillary effects, emissions taxes and trade (when countries play Nash) do not change the disappointing result of small stable climate coalitions, modesty, adaptation, altruism, reciprocity, support from outsiders, foresightedness may increase the stable climate coalition and may even result in the grand coalition.

The second strand our paper contributes to is the small literature on Kantian economics. Alger and Weibull (2013, 2016, 2020) have developed the framework of *homo moralis* preferences and proven that those preferences are evolutionary stable. As pointed out above, we follow Alger and Weibull, not least because their result that homo moralis preferences are evolutionary stable provides strong theoretical support for their approach. Daube and Ulph (2016) elaborate the consequences for environmental policy in a closed economy when individuals have preferences like those of *homo moralis*, but the behavior of their moral individuals deviates from that of Alger and Weibull and ours. Eichner and Pethig (2021) take up the *homo moralis* framework, but consider consumers with different degrees of morality and focus on decentralized climate policies without a climate coalition. Herweg and Schmidt (2022) consider moral consumers who suffer when deviating from a social norm, and investigate the impact of price versus quantity regulation on the consumers' moral behavior. A different concept of Kantian behavior underlies the Kant equilibrium approach of Laffont (1975) and Roemer (2010, 2015). Grafton et al. (2017) and Van Long (2020)

augment that approach by investigating the interaction of Kantian and Nashian agents in so-called Kant-Nash equilibria. The main message of this literature is that Kantian behavior improves public-good provision, is less harmful for the environment and may avoid the tragedy of the commons.

We are not aware of IEA games in the literature with countries exhibiting Kantian moral behavior of the Alger-Weibull type. The novelty of our paper is to model the moral behavior of governments with respect to both emissions and membership. In the subsequent Section 2 we discuss the IEA games sketched above in the order listed in Table 1.

2 IEA games with moral countries

As we argued in the introduction, countries may behave morally with respect to emissions and/or with respect to the decision to be a member of the fringe or coalition, and the moral behavior may be rigorous or less rigorous. In order to analyze such behavior when the formation of an IEA is at issue, we will set up a number of IEA games that differ with respect to the kind of morality affecting the countries' payoffs. In Section 2.1, we introduce the notation, the basic assumptions, and the game structure of the standard game of the early IEA literature without moral countries. That game will serve as a benchmark and a special case in subsequent IEA games with moral behavior. We deviate from the usual presentation of IEA games in the literature by explicitly modelling the countries' strategies to join or not to join the coalition. That procedure makes the behavioral assumptions in our IEA games precise and transparent, and therefore warrants the slightly more complex notation. For the benefit of exposition, we analyze and discuss different aspects of moral behavior successively. First, we deal with moral behavior with respect to emissions in its rigorous and moderate forms (Section 2.2), then we turn to moral behavior with respect to membership in its rigorous and moderate forms (Section 2.3) and finally combine both kinds of moral behavior to reach our ultimate goal which is the analysis of an IEA game with countries who act moderately morally on all dimensions (Section 2.4).

2.1 Notation, assumptions and the standard IEA game (Game $\boldsymbol{\sigma}$)

Throughout the paper, we consider two-stage IEA games in an international economy with the set $N = \{1, 2, \ldots, n\}$, $n > 3$, of identical countries. In the first stage, country i's strategy is the membership decision $s_i \in \{0, 1\}$, where $s_i = 1$ means "country i joins the agreement" and $s_i = 0$ means "country i remains an outsider". We refer to signatories as

coalition countries and to non-signatories as fringe countries. Any given strategy profile $\mathbf{s} := (s_1, \ldots, s_n) \in \{0,1\}^n$ uniquely defines the number of coalition countries $m = \sum_{j \in N} s_j$. The set of coalition countries is $C(\mathbf{s}) = \left\{ i \in N \,\middle|\, \mathbf{s} \in \{0,1\}^n \text{ and } s_i = 1 \right\}$, and the set of fringe countries is $F(\mathbf{s}) = N \setminus C(\mathbf{s})$. In the second stage, the strategy of country $i \in N$ is the emissions cap e_i of some global pollutant, such as a greenhouse gas. In the standard IEA game, country i's payoff is

$$W_\sigma^i(\mathbf{e}, \mathbf{s}) = B(e_i) - D\left(\sum_{j \in C(\mathbf{s})} e_j + \sum_{j \in F(\mathbf{s})} e_j \right), \tag{1}$$

where $\mathbf{e} = (e_1, \ldots, e_n) \in \mathbb{R}_+^n$. We refer to welfare $W_\sigma^i(\cdot)$ as materialistic welfare or σ-*welfare* and to the country with σ-welfare as materialistic country or *materialist*. $B(e_i)$ is the benefit, excluding any environmental damage, country i derives from production and consumption decisions that are made when the government has chosen the emission cap e_i. $D\left(\sum_{j \in C(\mathbf{s})} e_j + \sum_{j \in F(\mathbf{s})} e_j\right)$ is the environmental damage each country experiences. The functions B and D satisfy

$$B'(\cdot) > 0, B''(\cdot) < 0 \quad \text{and} \quad D(0) = 0, D'(\cdot) > 0, D'(\cdot) \geq 0 \tag{2a}$$

or alternatively

$$B'(e_i) = \left(\beta e_i - \frac{1}{2} e_i^2 \right) \quad \text{and} \quad D\left(\sum_{j \in N} e_j \right) = \delta \sum_{j \in N} e_j \quad \text{with } \beta > n\delta > 0. \tag{2b}$$

We will apply the parametric functional forms[6] (2b) only, if informative results cannot be derived with the more general functions B and D satisfying (2a).

The solution of the two-stage game is determined via backward induction as follows. When stage 2 is reached, all countries know the membership profile $\mathbf{s}$ that has been determined at stage 1. The coalition maximizes the aggregate welfare of its members

$$\sum_{h \in C(\mathbf{s})} W_\sigma^h(\mathbf{e}, \mathbf{s}) = \sum_{h \in C(\mathbf{s})} \left[B(e_h) - D\left(\sum_{j \in C(\mathbf{s})} e_j + \sum_{j \in F(\mathbf{s})} e_j \right) \right] \tag{3}$$

[6]The literature on the basic IEA game cited in the Introduction makes use of quadratic benefit functions, as in (2b), and either linear or quadratic damage functions. Assuming linear damage enables us to obtain analytical results on stable coalitions in Game $\varepsilon\mu\sigma$, whereas applying quadratic damages would yield informative results only via numerical simulations. The simulations we performed with quadratic damage functions suggest that our analytical results with linear damage to be presented below are robust to replacing linear by quadratic damage.

with respect to e_h for all $h \in C(\mathbf{s})$. The pertaining first-order conditions are

$$B'(e_i) - \sum_{j \in C(\mathbf{s})} D'\left(\sum_{j \in C(\mathbf{s})} e_j + \sum_{j \in F(\mathbf{s})} e_j \right) = 0 \qquad \forall\, i \in C(\mathbf{s}). \tag{4a}$$

Coalition countries internalize the damage they cause in their fellow coalition countries, but not the damage they cause in fringe countries. Insofar, the coalition exhibits self-interest vis-à-vis the fringe countries. It immediately follows from (3) and (4a) that in the grand coalition[7] ($\mathbf{s} = \mathbf{1}$) the full cooperation of all countries produces the socially optimal level of emissions, e^{SO}, and the corresponding socially optimal level of welfare, $w^{\mathrm{SO}} = W^i_\sigma(\mathbf{e}^{\mathrm{SO}}, \mathbf{1})$ in all countries.

For given $\sum_{j \in C(\mathbf{s})} e_j$ and $\sum_{j \in F(\mathbf{s}), j \neq i} e_j$, fringe country i maximizes (1) with respect to e_i. The corresponding first-order condition is

$$B'(e_i) - D'\left(\sum_{j \in C(\mathbf{s})} e_j + \sum_{j \in F(\mathbf{s})} e_j \right) = 0 \qquad \forall\, i \in F(\mathbf{s}). \tag{4b}$$

Fringe country i exhibits noncooperative and self-interested behavior, because it fails to internalize the damage it inflicts on all other countries.

The strategy profile $\mathbf{e}^* = \left(e_i^*, \mathbf{e}_{-i}^*\right) \in \mathbb{R}^n_+$ with $\mathbf{e}_{-i}^* := \left(e_1^*, \ldots, e_{i-1}^*, e_{i+1}^*, \ldots, e_n^*\right) \in \mathbb{R}^{n-1}_+$ constitutes a Nash equilibrium of the emissions subgame, if it satisfies

$$W^i_\sigma(e_i^*, \mathbf{e}_{-i}^*, \mathbf{s}) \;\geq W^i_\sigma(e_i, \mathbf{e}_{-i}^*, \mathbf{s}) \qquad \forall\, e_i \in \mathbb{R}_+,\ \forall\, i \in N. \tag{5}$$

We express the observation that the equilibrium payoffs $\mathbf{e}^*$ depend on the predetermined strategies $\mathbf{s}$ by the function[8]

$$e_i^* = E^i_\sigma(\mathbf{s}) \qquad \forall\, i \in N \tag{6}$$

and write the Nash equilibrium payoff as

$$W^i_\sigma\left[E^1_\sigma(\mathbf{s}), \ldots, E^n_\sigma(\mathbf{s}), \mathbf{s}\right] =: \tilde{W}^i_\sigma(\mathbf{s}). \tag{7}$$

Technically, the Nash equilibrium emissions $\mathbf{e}^*$ are determined by solving the equations (4a) and (4b).[9] For the boundary cases $\mathbf{s} = \mathbf{1}$ and $\mathbf{s} = \mathbf{0}$, the solutions are straightforward. If $\mathbf{s} = \mathbf{1}$ (grand coalition, $m = n$), it follows from (1) and (4a) that $\mathbf{e}^*$ is equal to the

[7] Observe that $\mathbf{1} := \{1, 1, \ldots, 1\} \in \mathbb{R}^n$ and $\mathbf{0} := \{0, 0, \ldots, 0\} \in \mathbb{R}^n$.

[8] To avoid clumsy notation we omit the subscript σ attached to the equilibrium strategies $e^*_{i\sigma}$ and $s^*_{i\sigma}$ and to the equilibrium welfare level $w^*_{i\sigma}$ whenever there is no risk of confusion.

[9] The equations (4a) and (4b) imply the countries' reaction functions.

socially optimal emissions $\mathbf{e}^{\mathrm{SO}}$. If $\mathbf{s} = \mathbf{0}$ (no coalition, $m = 0$), the equilibrium emissions are suboptimally large. Since the state of non-cooperation is often called business as usual we denote the corresponding equilibrium emissions $\mathbf{e}^* = \mathbf{e}^{\mathrm{BAU}}$. In all intermediate cases $1 < m < n$ the equilibrium emissions are suboptimally large, and the emissions of fringe countries are larger than those of coalition countries.

Next, we consider stage 1 of Game σ where the membership subgame is played. Country $\in N$ anticipates that its welfare will be $\tilde{W}^i_\sigma(\mathbf{s}) = \tilde{W}^i_\sigma(s_i, \mathbf{s}_{-i})$, defined in (7), when it chooses $s_i \in \{0, 1\}$ given the strategies $\mathbf{s}_{-i} = (s_1, \ldots, s_{i-1}, s_{i+1}, \ldots, s_n) \in \{0, 1\}^{n-1}$ of all other countries. Each country i maximizes $\tilde{W}^i_\sigma(s_i, \mathbf{s}_{-i})$ with respect to s_i, and the solution of the membership subgame is the Nash equilibrium strategy profile $\mathbf{s}^*$ satisfying

$$\tilde{W}^i_\sigma(s^*_i, \mathbf{s}^*_{-i}) \geq \tilde{W}^i_\sigma(s_i, \mathbf{s}^*_{-i}) \qquad \forall\, s_i \in \{0, 1\},\ \forall\, i \in N. \tag{8}$$

If $s^*_i = 1$ and country i switches to $s_i = 0$, it fails to be better off according to (8) (internal stability). If $s^*_i = 0$ and country i switches to $s_i = 1$, it also fails to be better off according to (8) (external stability). So, the equilibrium is stable in the sense that it exhibits internal and external stability. The equilibrium strategies $\mathbf{s}^*$ contain the information that the associated equilibrium coalition has $m^* = \sum_{j \in N} s^*_j$ members. The equilibrium strategies $\mathbf{s}^*$ inserted into the welfare $\tilde{W}^i_\sigma$ from (7) yields country i's equilibrium welfare level w^*_i.

Unfortunately, there is limited scope for informative results when the relatively general assumptions (2a) on the functions B and D are applied. In the IEA literature, Game σ (or an equivalent version of it) has been solved for less general parametric functions B and D such as (2b). We state the well-known result[10]

Proposition 1. *(Carraro and Siniscalco 1991, Hoel 1992)*
If the functions B and D satisfy the assumptions (2b), a coalition of three countries is stable in Game σ .

This disappointing result is due to the strong incentives of materialistic countries to free-ride by externalizing part of the damage they cause.

All IEA games with moral countries to be analyzed below will satisfy the structure of Game σ outlined in this section. The benefit function B and the damage function D will continue to be the only building blocks of the payoff functions. Moral behavior will be expressed by assuming (i) that countries deliberately make counterfactual assumptions on

[10] We will reproduce Proposition 1 as a side result in Proposition 5(iv).

the other countries' strategies and (ii) by blending different kinds of moral and/or non-moral behavior.

2.2 Moral behavior w.r.t. emissions (Games ε and $\varepsilon\sigma$)

Game ε. In the literature on IEAs, numerous analyses make use of materialistic welfare/payoff that is similar or equivalent to the σ-welfare (1). Here we argue that it may be more appropriate to assume that countries coping with global externality problems such as human-made climate change are guided by (their citizens') moral concerns. For such countries self-interested action is not "the right thing to do".[11] They realize (i) that their σ-welfare is low, if they choose a very high level of emissions and all other countries would do the same, and they realize (ii) that their σ-welfare would increase, if they choose a lower level of emissions and all other countries would do the same. Based on that consideration, they may conclude to act in the spirit of the Kantian categorical imperative which says that the right thing to do is choosing that level of emissions one would advocate all other countries choose as well. Following Alger and Weibull (2013, 2016, 2020), we capture the categorical imperative in our context by assigning to the moral country that σ-welfare, which it would enjoy on the counterfactual assumption that all other countries choose the same emissions. This contingent σ-welfare will be denoted *ε-welfare*, and countries with that kind of welfare are referred to as *emissions Kantians*. Formally, ε-welfare is defined as

$$W^i_\varepsilon(e_i) := W^i_\sigma(\mathbf{e}_{(i)}, \mathbf{s}) = B(e_i) - D(ne_i) \qquad \forall\, i \in N,\ \forall\, \mathbf{s} \in \{0,1\}^n, \tag{9}$$

where $\mathbf{e}_{(i)} := (e_i, \ldots, e_i) \in \mathbb{R}^n_+$. Let us suppose emissions Kantians are in a two-stage IEA game, called Game ε. In the second stage, they maximize (9) with respect to e_i, which yields the first-order condition

$$B'(e_i) - nD'(ne_i) = 0 \qquad \forall\, i \in N. \tag{10}$$

Since the payoff $W^i_\varepsilon(e_i)$, and therefore the first-order conditions (10), are independent of the emissions of all other countries, the emissions satisfying (10) are the equilibrium emissions $\mathbf{e}^*$. That is, the emissions subgame of Game ε has a Nash equilibrium in dominant strategies. It readily follows from the comparison of (10) with (4a) for $\mathbf{s} = \mathbf{1}$ that $\mathbf{e}^* = \mathbf{e}^{\text{SO}}$ such that the morally optimal equilibrium emissions are equal to the socially optimal emissions.

[11] Morally motivated tough climate policies are hardly conceivable without the morally motivated willingness of consumers and/or voters to bear the pertaining mitigation costs (Bernauer et al. 2016). In Eichner and Pethig (2021), the climate policy is moral to the extent that consumer-voters install governments that take action in accordance with the individuals' moral preferences.

The welfare $W_\varepsilon^i(e_i)$ is not only independent of $\mathbf{e}_{-i}$ but also independent of the membership strategies $\mathbf{s}$. So, emissions Kantians choose $\mathbf{e}^* = \mathbf{e}^{\text{SO}}$ no matter whether they are members of the coalition or not. Game ε is therefore a degenerate two-stage game - or as purists would argue - no game at all, because there is no strategic interdependence of players, when payoffs depend on own strategy but not on any other players' strategy. With this insight in mind, we keep using the term Game ε and summarize its solution in

Proposition 2. *(Game ε)*
If the functions B and D satisfy the assumptions (2a), the Nash equilibrium emissions of Game ε are socially optimal.

Game $\varepsilon\sigma$. Emissions Kantians exhibit moral rigor that can hardly be observed in individual, let alone in government behavior. It appears to be more appropriate to consider less rigorous moral behavior than that expressed in ε-welfare. Alger and Weibull (2013, 2016) provided theoretical support of this view by showing in the framework of evolutionary game theory that the evolution favors a particular class of preferences that attaches some weight to morality and some weight to self-interest. In Alger and Weibull (2020), they apply their concept of a moderately moral individual, they call *homo moralis*, in a simple closed economy with an emissions tax in which the formation of an IEA and the morality of governments is not an issue. The payoff function of their *homo moralis* is the weighted sum of the payoffs of a self-interested individual *(homo oeconomicus)* and a rigorous Kantian individual *(homo kantiensis)*. Here we take over Alger and Weibull's *homo moralis* concept. The analogue of their *homo oeconomicus* is our materialistic country of Game σ and the analogue of their *homo kantiensis* is our emissions Kantian of Game ε. Having in mind that we deviate from Alger and Weibull's setup by focusing on countries and coalition formation, we define the moral welfare, called $\varepsilon\sigma$-welfare, as

$$\begin{aligned} W_{\varepsilon\sigma}^i(\mathbf{e},\mathbf{s}) &:= \kappa_\varepsilon W_\varepsilon^i(e_i) + (1-\kappa_\varepsilon) W_\sigma^i(\mathbf{e},\mathbf{s}) \\ &= B(e_i) - \kappa_\varepsilon D(ne_i) - (1-\kappa_\varepsilon) D\left(\sum_{j\in C(\mathbf{s})} e_j + \sum_{j\in F(\mathbf{s})} e_j\right), \end{aligned} \tag{11}$$

that we will now consider as the payoff in an IEA game, called Game $\varepsilon\sigma$. In (11), $\kappa_\varepsilon \in [0,1]$ is a morality parameter we refer to as degree of ε-morality. Obviously, the extreme degrees of ε-morality, $\kappa_\varepsilon = 0$ and $\kappa_\varepsilon = 1$, are associated to the materialist and the emissions Kantian, respectively. So, the Game $\varepsilon\sigma$ contains the Games σ and ε as polar cases. Here we are interested in Game $\varepsilon\sigma$ with intermediate degrees of ε-morality that characterize countries, we call *emissions moralists*, whose moral action with respect to emissions is less rigorous

as that of emissions Kantians.[12] For $\kappa_\varepsilon \in]0,1[$, the equilibrium emissions of Game $\varepsilon\sigma$ are smaller than in Game σ ($e^*_{i\varepsilon\sigma} < e^*_{i\sigma}$) for all $i \in N$, larger than in Game ε ($e^*_{i\varepsilon\sigma} > e^*_{i\varepsilon} = e^{\mathrm{SO}}$) for all $i \in N$ and they tend to decrease in κ_ε toward $e^*_{i\varepsilon} = e^{\mathrm{SO}}$.

It may appear, at first glance, that there is a significant difference between σ- and $\varepsilon\sigma$-welfares. However, we can define

$$B_{\varepsilon\sigma}(e_i) := B(e_i) - \kappa_\varepsilon D(ne_i) \quad \text{and} \quad D_{\varepsilon\sigma}(\Sigma) := (1-\kappa_\varepsilon) D(\Sigma),$$

where $\Sigma := \sum_{j\in C(\mathbf{s})} e_j + \sum_{j\in F(\mathbf{s})} e_j$, and convert (11) into

$$W^i_{\varepsilon\sigma}(\mathbf{e},\mathbf{s}) = B_{\varepsilon\sigma}(e_i) - D_{\varepsilon\sigma}(\Sigma) =: V^i_\sigma(\mathbf{e},\mathbf{s};\kappa_\varepsilon). \tag{12}$$

It follows from (12) that the Games σ and $\sigma\varepsilon$ are isomorphic in the sense that for every Game $\varepsilon\sigma$ satisfying (2a) there exists a Game σ that is equivalent to that Game $\varepsilon\sigma$ and also satisfies (2a). So, it is possible to determine the solution of Game $\varepsilon\sigma$ with payoffs $W^i_{\varepsilon\sigma}(\mathbf{e},\mathbf{s})$ from (11) via determining the solution of Game σ with payoffs $V^i_\sigma(\mathbf{e},\mathbf{s};\kappa_\varepsilon)$ from (12).

Our observation above that specific information cannot be obtained on the solution of Game σ if the functions B and D satisfy (2a) extends to Game $\varepsilon\sigma$ due to (12). But in view of the isomorphism Proposition 1 applies which means that no Game $\varepsilon\sigma$ has a stable coalition with more than three members, if the functions B and D satisfy (2b). We summarize these findings in

Proposition 3. *(Game $\varepsilon\sigma$)*

(i) If the assumptions (2a) hold, Game $\varepsilon\sigma$ is isomorphic to Game σ.

(ii) If the less general assumptions (2b) hold, the stable coalition of Game $\varepsilon\sigma$ consists of three members.

Proposition 3(ii) squarely disappoints the expectation that stable coalitions in Game $\varepsilon\sigma$ are larger than in Game σ when (2b) is satisfied. They remain very small for all κ_ε lower than one. That is, when the share of σ-welfare in the $\varepsilon\sigma$-welfare is positive, that share dominates the outcome with respect to the stable coalition size. Beside that bad news, it is worth mentioning that in Game $\varepsilon\sigma$ the emissions of all countries are decreasing in κ_ε, and they reach the first-best level e^{SO}, when $\kappa_\varepsilon = 1$. Put differently, moderate morality with respect to emissions is good for curbing emissions, but bad in promoting large stable coalitions.

[12] In Daube and Ulph (2016), the moral utility also takes the form (11), but their individuals' behavior deviates from that of emissions moralists.

2.3 Moral behavior w.r.t. membership (Games μ and $\mu\sigma$)

Since Alger and Weibull (2020) do not focus on coalition formation, the level of emissions is their individuals' only moral concern. *Homo moralis* (who corresponds to our emissions moralist) takes full care of all moral concerns in their game. Our discussion above showed that coalition formation is obsolete or unnecessary in Game ε, because emissions Kantians implement the first best whether or not they are members of the coalition. But their indifference between being a member in the coalition or fringe turns out to be disadvantageous in the more realistic case of emissions moralists (Game $\varepsilon\sigma$), because then the materialist part of the countries' $\varepsilon\sigma$-welfare keeps the size of stable coalitions small.

When forming an IEA is at issue, the morality of emissions moralists (with $\varepsilon\sigma$-welfare) may be criticized as being onesided or biased, because their indifference with respect to membership implicitly sends the (wrong) signal that being indifferent between joining the coalition or not is the right thing to do. In other words, emissions moralists ignore that the choice of membership is a moral issue in its own right. Unbiased morality requires moral behavior on all dimensions, which in the present context calls for moral behavior with respect to both emissions *and* membership.

In the following, we will construct and analyze an IEA game with moral behavior on both dimensions, but before we do so, it is useful to specify and analyze membership morality in the absence of emissions morality. Like in Section 2.2, the first game with membership morality focuses on rigorous Kantian behavior (Game μ), and the second game on moderate Kantian behavior (Game $\mu\sigma$).

Game μ. A Kantian with respect to membership is a country whose moral welfare in the fringe [coalition] is equal to the σ-welfare it would enjoy on the counterfactual assumption that all other countries are in the fringe [coalition] as well. Formally, that moral welfare is defined as[13]

$$W^i_\mu(\mathbf{e}, s_i) := W^i_\sigma(\mathbf{e}, \mathbf{s}_{(i)}) = B(e_i) - D\left(\sum_{j\in H(\mathbf{s}_{(i)})} e_j\right) \tag{13}$$

where $\mathbf{s}_{(i)} := (s_i, \ldots, s_i) \in \{0,1\}^n$ and $H(\mathbf{s}_{(i)}) = \begin{cases} C\left(s_{(i)}\right) = C(\mathbf{1}) & \forall\, i \in C(\mathbf{s}), \\ F\left(s_{(i)}\right) = F(\mathbf{0}) & \forall\, i \in F(\mathbf{s}). \end{cases}$

We denote the welfare $W^i_\mu(\mathbf{e}, s_i)$ as *μ-welfare* and the countries with μ-welfare as *membership*

[13] Applying Alger and Weibull's Kantian behavior to our model means that country i counterfactually assumes that all other countries $j \neq i$ choose its strategy. In Game ε country i assumes $\mathbf{e} = \mathbf{e}_{(i)}$ and in Game μ country i assumes $\mathbf{s} = \mathbf{s}_{(i)}$.

Kantians. The μ-welfare (13) is defined in analogy to the ε-welfare and captures well, as we believe, the spirit of Kant's categorical imperative applied to the choice of membership.

To solve Game μ we assume that stage 2 of that game is reached, where the decisions to join the coalition has already been taken. For given memberships $\mathbf{s}$ the coalition country maximizes $\sum_{j\in C(\mathbf{s}_{(i)})} W_\mu^j(\mathbf{e}, s_i)$ with respect to e_i and the fringe country $i \in F(\mathbf{s})$ maximizes $W_\mu^i(\mathbf{e}, s_i)$ with respect to e_i. The pertaining first-order conditions yield

$$B'(e_i) - \sum_{j\in C(\mathbf{1})} D'\left(\sum_{j\in C(\mathbf{1})} e_j\right) = 0 \qquad \forall\, i \in C(\mathbf{s}), \tag{14a}$$

$$B'(e_i) - D'\left(\sum_{j\in F(\mathbf{0})} e_j\right) = 0 \qquad \forall\, i \in F(\mathbf{s}). \tag{14b}$$

Equation (14a) is equal to equation (4a) for $\mathbf{s} = \mathbf{1}$ implying that if a membership Kantian is a member of the really existing coalition, i.e. of the coalition that consists of all countries i with $s_i = 1$ in the given membership profile $\mathbf{s}$, it plays the dominant strategy $\tilde{E}_\mu^i(\mathbf{s}_{(i)}) = \tilde{E}_\mu^i(\mathbf{1})$ equal to the socially optimal level e^{SO} of emissions. Conversely, equation (14b) equals equation (4b) for $\mathbf{s} = \mathbf{0}$ implying that membership Kantians in the really existing fringe play the dominant strategy $\tilde{E}_\mu^i(\mathbf{s}_{(i)}) = \tilde{E}_\mu^i(\mathbf{0})$ that is equal to the high buisiness-as-usual level e^{BAU} of emissions. So, the Nash equilibrium of the emissions subgame of Game μ is characterized by

- the emissions $\tilde{E}_\mu^i(\mathbf{1}) = e^{\mathrm{SO}}$ and the welfare $\tilde{W}_\mu^i(\mathbf{1}) = w^{\mathrm{SO}}$ for all $i \in C(\mathbf{s})$ and
- the emissions $\tilde{E}_\mu^i(\mathbf{0}) = e^{\mathrm{BAU}}$ and the welfare $\tilde{W}_\mu^i(\mathbf{0}) = w^{\mathrm{BAU}}$ for all $i \in F(\mathbf{s})$.

Given these observations and the obvious fact that $w^{\mathrm{SO}} > w^{\mathrm{BAU}}$, the solution of the membership subgame at stage 1 is straightforward. If $\mathbf{s}$ satisfies $2 < m < n$, joining the fringe would make coalition countries worse off. So coalitions are internally stable. But since fringe countries are better off when joining the coalition (external instability) we obtain

Proposition 4. *(Game μ)*
If the functions B and D satisfy (2a), the only stable coalition of Game μ is the grand coalition which implements the social optimum.

Game $\mu\sigma$. Our assessment that the rigorous morality of emissions Kantians in Game ε can hardly be observed in practice also applies to membership Kantians in Game μ. So in analogy to the concept of the degree of ε-morality, κ_ε, we introduce the degree of μ-morality, $\kappa_\mu \in [0, 1]$, and define the $\mu\sigma$-welfare as the welfare that is equal to the sum of the μ-welfare

with weight κ_μ and the σ-welfare with weight $(1-\kappa_\mu)$. In formal terms,

$$W^i_{\mu\sigma}(\mathbf{e},\mathbf{s}) := \kappa_\mu W^i_\mu(\mathbf{e},s_i) + (1-\kappa_\mu) W^i_\sigma(\mathbf{e},\mathbf{s}). \tag{15}$$

Invoking the definitions of $W^i_\mu(\mathbf{e},s_i)$ and $W^i_\sigma(\mathbf{e},\mathbf{s})$, we rewrite (15) as

$$W^i_{\mu\sigma}(\mathbf{e},\mathbf{s}) = B(e_i) - \kappa_\mu D\Bigg(\sum_{j\in H(\mathbf{s}_{(i)})} e_j\Bigg) - (1-\kappa_\mu) D\Bigg(\sum_{j\in C(\mathbf{s})} e_j + \sum_{j\in F(\mathbf{s})} e_j\Bigg), \tag{16}$$

where $H(\mathbf{s}_{(i)}) = \begin{cases} C\left(s_{(i)}\right) = C(\mathbf{1}) & \forall\, i \in C(\mathbf{s}), \\ F\left(s_{(i)}\right) = F(\mathbf{0}) & \forall\, i \in F(\mathbf{s}). \end{cases}$

The first-order conditions of maximizing $\sum_{j\in C(\mathbf{s})} W^j_{\mu\sigma}(\mathbf{e},s_i)$ with respect to e_i for $i \in C(\mathbf{s})$ and maximizing $W^i_{\mu\sigma}(\mathbf{e},s_i)$ with respect to e_i for $i \in F(\mathbf{s})$ are

$$B'(e_i) - \kappa_\mu \sum_{j\in C(\mathbf{1})} D'\Big(\sum_{j\in C(\mathbf{1})} e_j\Big) - (1-\kappa_\mu) \sum_{j\in C(\mathbf{s})} D'\Big(\sum_{j\in C(\mathbf{s})} e_j + \sum_{j\in F(\mathbf{s})} e_j\Big) = 0 \quad \forall\, i \in C(\mathbf{s}), \tag{17a}$$

$$B'(e_i) - \kappa_\mu D'\Big(\sum_{j\in F(\mathbf{0})} e_j\Big) - (1-\kappa_\mu) D'\Big(\sum_{j\in C(\mathbf{s})} e_j + \sum_{j\in F(\mathbf{s})} e_j\Big) = 0 \quad \forall\, i \in F(\mathbf{s}). \tag{17b}$$

As outlined in Section 2.1 for Game σ, the solution of the equations (17a) and (17b) for emissions yields the Nash equilibrium emissions $e^*_i = E^i_{\mu\sigma}(\mathbf{s})$ and the Nash equilibrium payoffs $W^i_{\mu\sigma}\left[E^1_{\mu\sigma}(\mathbf{s}),\ldots,E^n_{\mu\sigma}(\mathbf{s}),\mathbf{s}\right] =: \tilde{W}^i_{\mu\sigma}(\mathbf{s})$. So the Nash equilibrium of the membership subgame is a membership profile $\mathbf{s}^*$ satisfying

$$\tilde{W}^i_{\mu\sigma}(s^*_i,\mathbf{s}^*_{-i}) \geq \tilde{W}^i_{\mu\sigma}(s_i,\mathbf{s}^*_{-i}) \qquad \forall\, s_i \in \{0,1\},\ \forall\, i \in N. \tag{18}$$

In order to characterize that equilibrium, it is analytically convenient to apply here and in the remainder of the paper a procedure that is widely used in the literature. We take advantage of the observation that in equilibrium all members in their group choose the same emissions. So, we now write (without loss of generality) $e_i = e_c$, if $i \in C(\mathbf{s})$, and $e_i = e_f$, if $i \in F$, even before the equilibrium is reached. Moreover, closer inspection shows that

$$E^i_{\mu\sigma}(\mathbf{s}) = E^i_{\mu\sigma}(\mathbf{s}') \text{ and } \tilde{W}^i_{\mu\sigma}(\mathbf{s}) = \tilde{W}^i_{\mu\sigma}(\mathbf{s}') \quad \Longleftrightarrow \quad \sum_{j\in N} s_j = \sum_{j\in N} s'_j =: m,$$

which allows us to redefine the functions $E^i_{\mu\sigma}$ and $W^i_{\mu\sigma}$ as

$$E^i_{\mu\sigma}(\mathbf{s}) = \begin{cases} \hat{E}^c_{\mu\sigma}(m) & \text{if } i \in C(\mathbf{s}), \\ \hat{E}^f_{\mu\sigma}(m) & \text{if } i \in F(\mathbf{s}), \end{cases} \quad \text{and} \quad \tilde{W}^i_{\mu\sigma}(\mathbf{s}) = \begin{cases} \hat{W}^c_{\mu\sigma}(m) & \text{if } i \in C(\mathbf{s}), \\ \hat{W}^f_{\mu\sigma}(m) & \text{if } i \in F(\mathbf{s}). \end{cases} \tag{19}$$

That transformation greatly simplifies the analysis, because it allows to determine an equilibrium strategy profile $\mathbf{s}^*$ indirectly via determining the size m^* of an equilibrium coalition.

Specifically, the Nash equilibrium strategy profile s^* defined in (18) is equivalent to the size of the stable coalition $m^* = \sum_{j \in N} s_j^*$ if m^* satisfies

$$S_{\mu\sigma}(m^*) \geq 0 \text{ (internal stability)} \quad \text{and} \quad S_{\mu\sigma}(m^*+1) < 0 \text{ (external stability)}, \tag{20}$$

where $S_{\mu\sigma}(m) := \hat{W}^c_{\mu\sigma}(m) - \hat{W}^f_{\mu\sigma}(m-1)$. A coalition of size m^* is stable if no coalition country has an incentive to leave the coalition (internal stability) and no fringe country to join the coalition (external stability). We prove in the Appendix

Proposition 5. *(Game $\mu\sigma$)*

(i) Suppose the functions B and D satisfy (2a).[14] *Subject to minor restrictions, there exists $\bar{\kappa}_\mu \in]0,1[$ such that*

$$\hat{W}^c_{\mu\sigma}(m;\kappa_\mu) \lesseqgtr \hat{W}^f_{\mu\sigma}(m;\kappa_\mu) \quad \Longleftrightarrow \quad \kappa_\mu \lesseqgtr \bar{\kappa}_\mu \quad \forall\, m \in \{2,\ldots,n-1\}. \tag{21}$$

(ii) Suppose the functions B and D satisfy (2b). The size of the stable coalition is $m^* = 3$, if $\kappa_\mu = 0$; there exists $\tilde{\kappa}_\mu < 1$ such that the stable coalition size m^* is strictly monotone increasing in κ_μ on $[0, \tilde{\kappa}_\mu[$; it is equal to $m^* = n$ for all $\kappa_\mu \in [\tilde{\kappa}_\mu, 1]$.[15]

To understand Proposition 5(i), recall from our discussion of the Games σ and μ that for given m the extreme cases $\kappa_\mu = 0$ and $\kappa_\mu = 1$ are characterized by $\hat{W}^c_{\mu\sigma}(m;0) < \hat{W}^f_{\mu\sigma}(m;0)$ and $\hat{W}^c_{\mu\sigma}(m;1) > \hat{W}^f_{\mu\sigma}(m;1)$, respectively. So, Proposition 5(i) establishes that in the transition from $\kappa_\mu = 0$ to $\kappa_\mu = 1$ the welfares curves are continuous and intersect only once. The advantage in (moral) welfare the fringe countries have in case of $\kappa_\mu = 0$ is decreasing in κ_μ and turns into a disadvantage for sufficiently large κ_μ.

Some comments are in order on Proposition 5(iia). For the stability analysis (20) it is relevant how the difference $\hat{W}^f_{\mu\sigma}(m;\kappa_\mu) - \hat{W}^c_{\mu\sigma}(m;\kappa_\mu)$ depends on m for alternatively given degrees of μ-morality. Making use of the parametric functions (2b), we illustrate the difference $\hat{W}^f_{\mu\sigma}(m;\kappa_\mu) - \hat{W}^c_{\mu\sigma}(m;\kappa_\mu)$ in the graphs of Figure 1. These graphs are based on the numerical specification of parameters $\beta = 200$, $\delta = 1$, $n = 100$, and they differ in that $\kappa_\mu = 0.1$ is raised to $\kappa_\mu = 0.4$.

[14] The notation $\hat{W}^q_{\mu\sigma}(m;\kappa_\mu)$ for $q = c, f$ makes visible that the function $\hat{W}^q_{\mu\sigma}$ also depends on the parameter κ_μ.

[15] The result in Proposition 5(ii) that $m^* = 3$, if $\kappa_\mu = 0$, proves Proposition 1. The case $\kappa_\mu = 1$ means we deal with Game μ for which the size of the stable coalition is $m^* = n$ (Proposition 4).

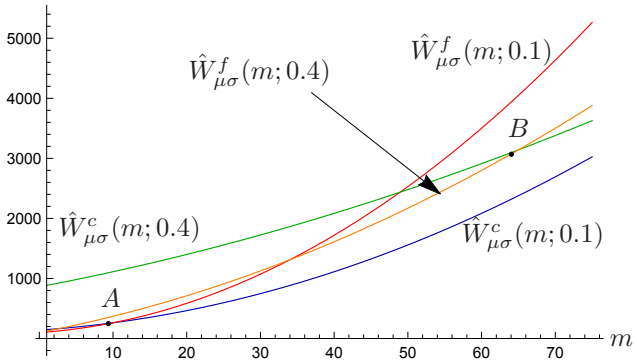

Figure 1: Dependence on m of the equilibrium $\mu\sigma$-welfare when κ_μ is raised from $\kappa_\mu = 0.1$ to $\kappa_\mu = 0.4$

The curves relating to the same κ_μ satisfy $\hat{W}^c_{\mu\sigma}(1;\kappa_\mu) > \hat{W}^f_{\mu\sigma}(1;\kappa_\mu)$ and $\hat{W}^c_{\mu\sigma}(n-1;\kappa_\mu) < \hat{W}^f_{\mu\sigma}(n-1;\kappa_\mu)$. Both countries' moral welfare is increasing in m, but the fringe country's curve is steeper such that the curves intersect. In Figure 1 the points A and B satisfy $\hat{W}^c(m;0.1) = \hat{W}^f(m;0.1)$ and $\hat{W}^c(m;0.4) = \hat{W}^f(m;0.4)$, respectively. So, the m-coordinate of point B is is larger than that of point A. It will turn out to be important that the intersection point of the curves of coalition and fringe countries (for given κ_μ) shifts to the right when κ_μ increases.

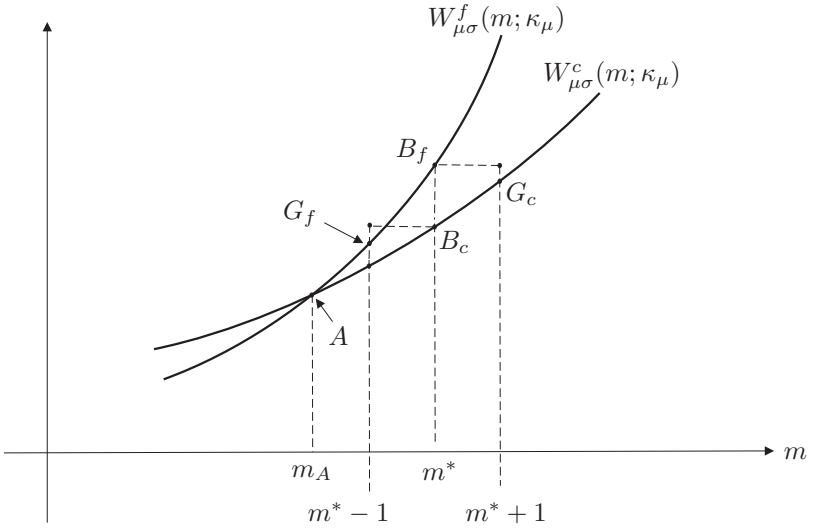

Figure 2: Illustration of a stable coalition

How Figure 1 relates to the stability of coalitions can be conveniently discussed in the free-hand Figure 2. That figure exhibits a section of the graphs in Figure 1 for some given κ_μ. We will demonstrate that if there is a stable coalition size m^* when the curvature of the

welfare curves is as in Figure 1, then m^* must be larger than m_A and close to m_A. We have drawn m^* in Figure 2 such that it is in fact the size of a stable coalition. To see that observe first that if m^* prevails, the coalition countries' welfare is $\hat{W}^c_{\mu\sigma}(m^*;\kappa_\mu)$ in point B_c and the fringe countries' welfare is $\hat{W}^f_{\mu\sigma}(m^*;\kappa_\mu)$ in point B_f. When leaving the coalition, the coalition country moves from point B_c to point G_f where its welfare $\hat{W}^f_{\mu\sigma}(m^*-1;\kappa_\mu)$ is smaller than $\hat{W}^c_{\mu\sigma}(m^*;\kappa_\mu)$ (internal stability). When leaving the fringe, the fringe country moves from point B_f to point G_c where its welfare $\hat{W}^c_{\mu\sigma}(m^*+1;\kappa_\mu)$ is smaller than $\hat{W}^f_{\mu\sigma}(m^*;\kappa_\mu)$ (external stability). That proves stability. Closer inspection of the curvature of the graphs in Figure 1 shows (i) that all coalitions of size $m \leq m_A$ are externally unstable and (ii) that all coalitions of size m are internally unstable, if $m - m_A$ is positive and sufficiently large. It follows that if there exists a stable coalition in Game $\mu\sigma$ for the parameters assumed in Figure 1, then (i) its size must be positioned as drawn in Figure 2 and (ii) in view of Figure 1 m^* must increase if $\kappa_\mu = 0.1$ is raised to $\kappa_\mu = 0.4$.

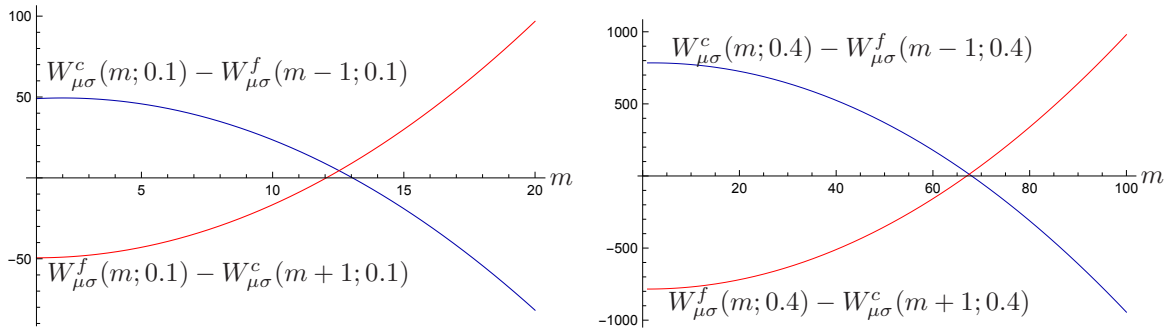

Figure 3: Stable coalitions, when parameters are as in Figure 1

The existence of such stable coalitions is confirmed in Figure 3, where the stability procedure (20) is applied to the Game $\mu\sigma$ with the parameters of Figure 1. By definition of the graphs, a coalition of size m is stable if and only if m is in that (small) interval in which both graphs are positive. Inspection of Figure 3 shows that we have $m^* = 13$ in the left panel of Figure 3 and $m^* = 68$ in the right panel of Figure 3. The reason why the increase in κ_μ results in an increase in m^* is readily seen in Figure 1. Knowing that the stable coalitions are located to the right of and close to the intersection points in Figure 1, the question needs to be answered, why the intersection point of the welfare curves for $\kappa_\mu = 0.4$ is located to the right of that for $\kappa_\mu = 0.1$. Inspection of Figure 1 shows that the increase in κ_μ shifts upward the welfare curves of both coalition and fringe countries. However, that shift is stronger for coalition than for fringe countries such that the intersection point shifts

to the right and with it the size of the stable coalition.

This result is generalized in Proposition 5(ii) and verified in Figure 4 where we depict the graph of the function[16] $\mathring{M}(\kappa_\mu, n)$. That graph maps all $\kappa_\mu \in [0,1]$ and all n into m^* and demonstrates that the μ-morality is a powerful means to secure large stable coalitions. The stability of the grand coalition can be achieved even for degrees of μ-morality smaller than $\kappa_\mu = 1$. For $n = 100$ the grand coalition is stable for all $\kappa_\mu > \tilde{\kappa}_\mu := 0.49745$.

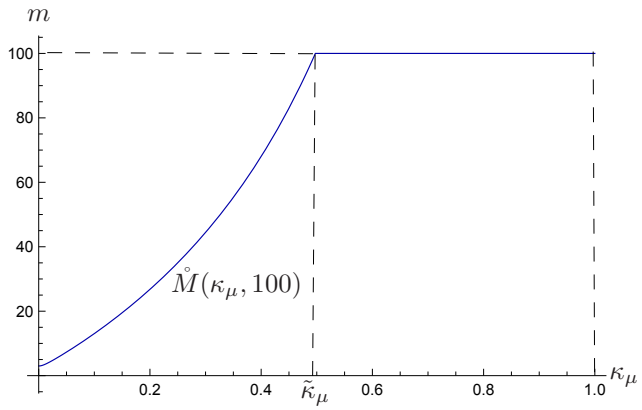

Figure 4: Size of stable coalitions dependent on κ_μ

2.4 Moral behavior w.r.t. both emissions and membership (Game $\varepsilon\mu\sigma$)

We have argued above that the Kantian moral rigor in the Games ε and μ is unrealistic and have focused on moderate Kantian morality in the Games $\varepsilon\sigma$ and $\mu\sigma$ with degrees of κ_ε- and κ_μ-morality smaller than one. But assuming that countries exhibit moderate Kantian behavior either with respect to emissions only or with respect to membership only is still unsatisfactory because such biased moral behavior appears to be implausible. To model moral behavior 'on both dimensions', we now define a Game $\varepsilon\mu\sigma$ with payoffs, denoted $\varepsilon\mu\sigma$-welfares, that are a convex combination of $\varepsilon\sigma$-welfare and $\mu\sigma$-welfare analogous to our procedure in the Games $\varepsilon\sigma$ and $\mu\sigma$, where we introduced a convex combination between ε-welfare or μ-welfare on the one hand and σ-welfare on the other. Formally, country i's $\varepsilon\mu\sigma$ -welfare is given by

$$W^i_{\varepsilon\mu\sigma}(\cdot) := \alpha W^i_{\mu\sigma}(\cdot) + (1-\alpha) W^i_{\varepsilon\sigma}(\cdot) \qquad \forall\ i \in N, \tag{22}$$

[16]The function $\mathring{M}$ is derived in the Appendix. It is interesting to observe that $\mathring{M}$ does not depend on the parameters β and δ.

where $\kappa_\varepsilon \in [0,1]$, $\kappa_\mu \in [0,1]$ and where the new parameter $\alpha \in [0,1]$ is the moral weight of $\mu\sigma$-welfare. Invoking the definitions of $W^i_{\varepsilon\sigma}(\cdot)$ and $W^i_{\mu\sigma}(\cdot)$ we rewrite (22) as

$$W^i_{\varepsilon\mu\sigma}(\cdot) = \alpha\kappa_\mu W^q_\mu(\cdot) + (1-\alpha)\kappa_\varepsilon W^i_\varepsilon(\cdot) + [(1-\alpha)(1-\kappa_\varepsilon) + \alpha(1-\kappa_\mu)]\, W^i_\sigma(\cdot). \tag{23}$$

One may argue that the analysis should be restricted to the non-biased morality parameters $\alpha = 1/2$ and $\kappa_\mu = \kappa_\varepsilon = \kappa \in]0,1[$, because deviations from these parameters appear to be implausible. However, from a theoretical point of view it is interesting to characterize the outcome of all possible combinations of the morality parameters α, κ_ε and κ_μ. After all, that procedure does not exclude the case of non-biased morality.

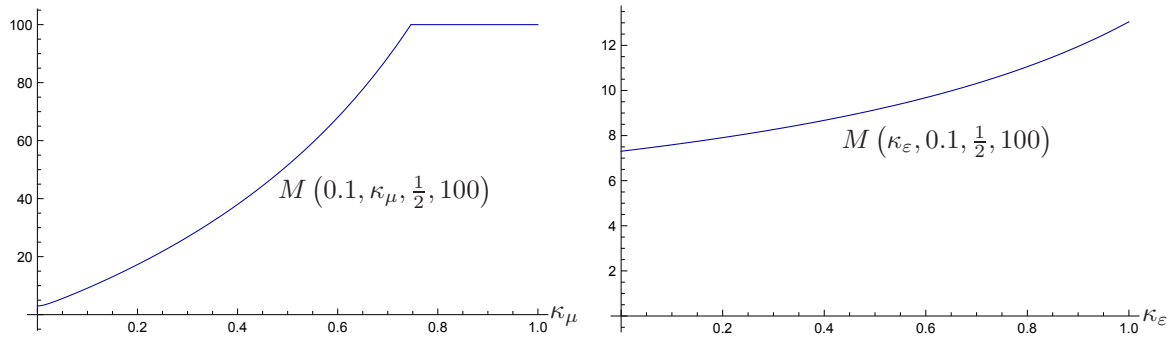

Figure 5: Stable coalitions

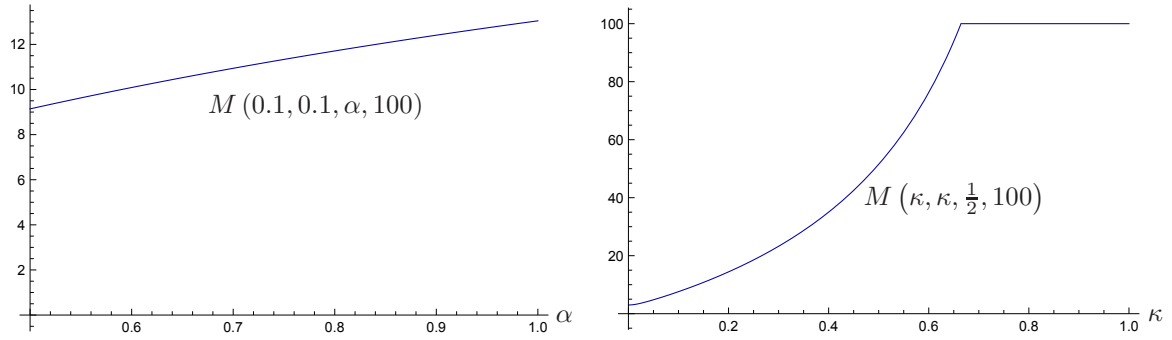

Figure 6: Stable coalitions

Due to the complexity of $\varepsilon\mu\sigma$-welfare (22), informative results cannot be obtained unless we reduce generality by applying the parametric functions B and D of (2b). In the Appendix, we derive a function of the type[17]

$$m^* = M(\alpha, \kappa_\varepsilon, \kappa_\mu, n). \tag{24}$$

[17]Observe that the function M from (24) and $\mathring{M}$ from Figure 4 are related by $M(0, \kappa_\varepsilon, \kappa_\mu, n) \equiv \mathring{M}(\kappa_\mu, n)$.

That function provides the information that in Game $\varepsilon\mu\sigma$ with n countries and the morality parameters $(\alpha, \kappa_\varepsilon, \kappa_\mu) \in [0,1]^3$ the stable coalition has $M(\alpha, \kappa_\varepsilon, \kappa_\mu, n)$ members. Thus, (24) fully characterizes the stable coalitions in Game $\varepsilon\mu\sigma$. We begin our discussion of (24) with the two graphs in Figure 5 and the graph in the left panel of Figure 6 showing how the size of the stable coalition responds to a partial variation of one of the parameters κ_ε, κ_μ and α. The common feature is that the size of the stable coalition is increasing in each of these parameters. The steepness of the curves and whether the grand coalition is reached depends on the values at which the other parameters are kept constant. More information on that will be provided below.

The graph plotted in the right panel of Figure 6 relates to the interesting case of a fully symmetric (or unbiased) general moralist characterized by $\alpha = 0.5$ and $\kappa_\varepsilon = \kappa_\mu = \kappa \in [0,1]$. If $\kappa_\varepsilon = \kappa_\mu$ and both degrees of morality are raised by the same amount, the size of the stable coalition increases and the grand coalition is reached at degrees of morality well below one.

To understand why the curves presented in the Figures 5 and 6 are upward sloping, we computed the effects of a non-marginal change in the parameters κ_ε, κ_μ, α and κ, respectively, and plotted the results of each of these parameter changes in two figures analogous to the Figures 1 and 3 in Game $\mu\sigma$. These new figures turned out to be the same as the Figures 1 and 3, in qualitative terms. It suffices, therefore, to point out that an increase in each of the morality parameter α, κ_ε and κ_μ causes an upward shift of the welfare curves of both coalition and fringe countries, as it does in Figure 1, and that shift is larger for coalition than for fringe countries.[18]

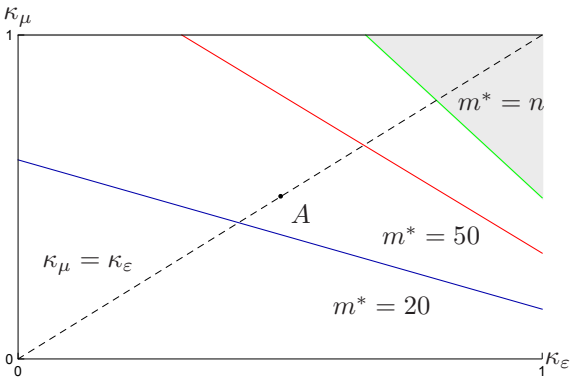

Figure 7: Characterization of stable coalitions for $n = 100$ and $\alpha = 0.25$

Although the Figures 5 and 6 yield interesting results it is unclear how general that

[18] An exception is the unexpected result of the partial variation of κ_ε which is further explained in the Figures 11 and 12 below.

information is. An answer to that question is given in the Figures 7 - 9 that are graphical presentations of the function (24) generated with mathematica. For $\alpha = 0.25$, $\alpha = 0.5$ and $\alpha = 0.75$, respectively, the rectangles in these figures represent the set $T(\alpha) = \left\{ (\kappa_\varepsilon, \kappa_\mu) \;\middle|\; (\alpha, \kappa_\varepsilon, \kappa_\mu) \in \mathcal{T} \right\}$, where $\mathcal{T}$ is the set of all feasible triples $(\alpha, \kappa_\varepsilon, \kappa_\mu)$. The stable coalition sizes m^* assigned to the triples $(\alpha, \kappa_\varepsilon, \kappa_\mu)$ are indirectly accounted for in the Figures 7 - 9 by the m^*-isoquants for $m^* = 20$ and $m^* = 50$ and by the sets $\bar{T}(\alpha) \subset T(\alpha)$ of all $(\kappa_\varepsilon, \kappa_\mu)$ for which $m^* = n = 100$ and α is given. $\bar{T}(\alpha)$ is represented in the Figures 7 - 9 by the shaded areas. The non-shaded areas in these figures correspond to the sets $T(\alpha) \setminus \bar{T}(\alpha)$ of tuples $(\kappa_\varepsilon, \kappa_\mu)$ that yield stable coalitions smaller than n. The three-dimensional shaded area in Figure 10 is the lower bound of all triples $(\alpha, \kappa_\varepsilon, \kappa_\mu)$ for which $m^* = n$.

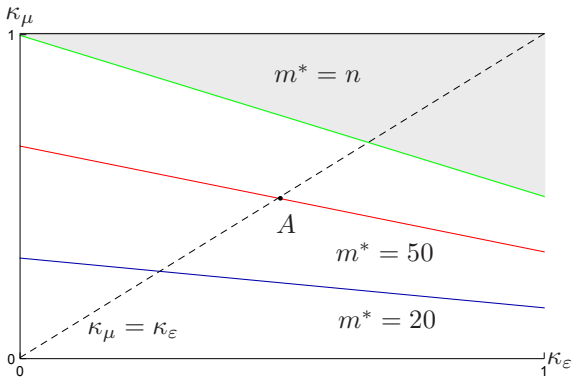

Figure 8: Characterization of stable coalitions for $n = 100$ and $\alpha = 0.5$

It is straightforward to see that the graphs in the Figures 5 and 6 which relate to the partial variations in κ_ε, κ_μ and κ depict m^* for the tuples along a horizontal, vertical or diagonal cut through the rectangles in the Figures 7 - 9. That m^* is increasing in α (left panel of Figure 6) is demonstrated indirectly in the Figures 7 - 9 as follows. The point A is placed such that its coordinates $(\kappa_\varepsilon, \kappa_\mu)$ are the same in each of the figures. One can readily see that the size of the stable coalition is $m^* < 50$ if $\alpha = 0.25$ in Figure 7, $m^* < 50$ if $\alpha = 0.5$ in Figure 8 and $m^* > 50$ if $\alpha = 0.75$ in Figure 9.

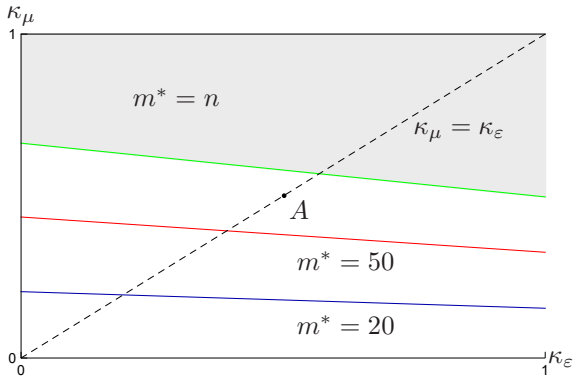

Figure 9: Characterization of stable coalitions for $n = 100$ and $\alpha = 0.75$

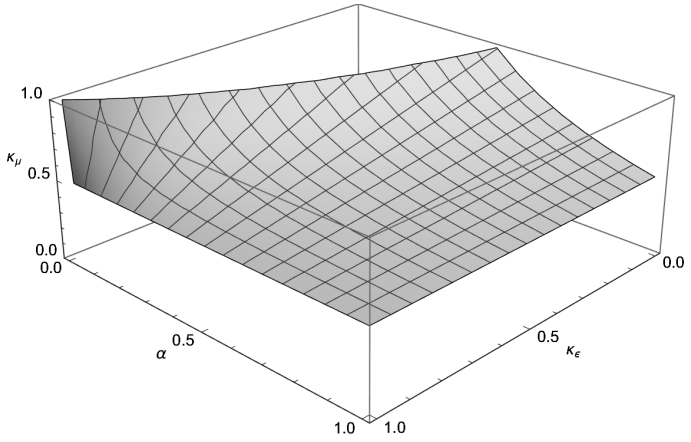

Figure 10: Characterization of the stable grand coalition for $n = 100$

In the Appendix we prove the main properties of the solution of Game $\varepsilon\mu\sigma$ which are summarized in

Proposition 6. *(Game $\varepsilon\mu\sigma$)*
Suppose the functions B and D satisfy (2b).

(i) If $(\alpha, \kappa_\varepsilon, \kappa_\mu) \in \mathcal{T} \setminus \bar{\mathcal{T}}$, there exists a function $m^* = M(\alpha, \kappa_\varepsilon, \kappa_\mu, n)$ with the properties

$$\frac{\partial m^*}{\partial \kappa_\varepsilon} = M_{\kappa_\varepsilon} > 0, \frac{\partial m^*}{\partial \kappa_\mu} = M_{\kappa_\mu} > 0, \frac{\partial m^*}{\partial \alpha} = M_{\kappa_\alpha} > 0, \frac{\partial M\left(\frac{1}{2}, \kappa, \kappa, n\right)}{\partial \kappa} > 0. \tag{25}$$

(ii) The grand coalition is stable, if and only if $(\alpha, \kappa_\mu, \kappa_\varepsilon) \in \bar{\mathcal{T}}$. The set $\bar{\mathcal{T}}$ is non-empty and a proper subset of the set of feasible economies.

(a) For all $(\alpha, \tilde{\kappa}_\varepsilon, \kappa_\mu)$ satisfying $S_{\varepsilon\mu\sigma}(m^* = n; \alpha, \tilde{\kappa}_\varepsilon, \kappa_\mu) = 0$, it holds

$$(\alpha, \kappa_\varepsilon, \kappa_\mu) \left\{ \begin{array}{c} \in \bar{\mathcal{T}} \\ \notin \bar{\mathcal{T}} \end{array} \right\} \iff \kappa_\varepsilon \left\{ \begin{array}{c} \geq \\ < \end{array} \right\} \tilde{\kappa}_\varepsilon.$$

(b) For all $(\alpha, \kappa_\varepsilon, \tilde{\kappa}_\mu)$ *satisfying* $S_{\varepsilon\mu\sigma}(m^* = n, \alpha, \kappa_\varepsilon, \tilde{\kappa}_\mu) = 0$*, it holds*

$$(\alpha, \kappa_\varepsilon, \kappa_\mu) \left\{ \begin{array}{c} \in \bar{\mathcal{T}} \\ \notin \bar{\mathcal{T}} \end{array} \right\} \iff \kappa_\mu \left\{ \begin{array}{c} \geq \\ < \end{array} \right\} \tilde{\kappa}_\mu.$$

(c) For all $(\tilde{\alpha}, \kappa_\varepsilon, \kappa_\mu)$ *satisfying* $S_{\varepsilon\mu\sigma}(m^* = n; \tilde{\alpha}, \kappa_\varepsilon, \kappa_\mu) = 0$*, it holds*

$$(\alpha, \kappa_\varepsilon, \kappa_\mu) \left\{ \begin{array}{c} \in \bar{\mathcal{T}} \\ \notin \bar{\mathcal{T}} \end{array} \right\} \iff \alpha \left\{ \begin{array}{c} \geq \\ < \end{array} \right\} \tilde{\alpha}.$$

(d) For all $\left(\alpha = \frac{1}{2}, \kappa_\varepsilon = \tilde{\kappa}, \kappa_\mu = \tilde{\kappa}\right)$ *satisfying* $S_{\varepsilon\mu\sigma}(m^* = n; \alpha, \tilde{\kappa}, \tilde{\kappa}) = 0$*, it holds*

$$\left(\alpha = \tfrac{1}{2}, \kappa_\varepsilon = \kappa, \kappa_\mu = \kappa\right) \left\{ \begin{array}{c} \in \bar{\mathcal{T}} \\ \notin \bar{\mathcal{T}} \end{array} \right\} \iff \kappa \left\{ \begin{array}{c} \geq \\ < \end{array} \right\} \tilde{\kappa}.$$

While Proposition 6(i) is self-explanatory, some remarks are necessary for Proposition 6(ii). Recall that $S_{\varepsilon\mu\sigma}(m; \alpha, \kappa_\varepsilon, \kappa_\mu) = W^c_{\varepsilon\mu\sigma}(m; \alpha, \kappa_\varepsilon, \kappa_\mu) - W^f_{\varepsilon\mu\sigma}(m-1; \alpha, \kappa_\varepsilon, \kappa_\mu)$ is the stability function of Game $\varepsilon\mu\sigma$ and observe that in case of the grand coalition ($m^* = n$) only the external stability is relevant. $S_{\varepsilon\mu\sigma}(m^* = n; \cdot) = 0$ holds on the $m^* = n$-isoquants (green lines[19]) in Figure 7 - 9 and on the grey-shaded area in Figure 10. Proposition 6(iia) then states that starting from a point on the $m^* = n$-isoquants with some $\tilde{\kappa}_\varepsilon$, leaving all other parameters unchanged, the grand coalition is stable for all $\kappa_\varepsilon \geq \tilde{\kappa}_\varepsilon$. Analogue arguments apply to Proposition 6(iib)-(6iid).

The principal message is that in Game $\varepsilon\mu\sigma$ the stable coalition is the larger and the grand coalition is the more likely stable

(a) the larger the μ-morality,

(b) the larger the ε-morality,

(c) the larger α.

Since an increase of κ_μ already increased the stable coalition and made the stability of the grand coalition more likely in Game $\mu\sigma$ (Proposition 5(ii)), whereas an increase of κ_ε does not increase the stable coalition in Game $\varepsilon\sigma$ (Proposition 3(ii)) the messages (a) and (c) are as expected, but the message (b) is surprising, if not counterintuitive. To understand the driving force for the effects of ε-morality, we keep α and κ_μ constant and increase κ_ε. From the first-order conditions in Game $\varepsilon\mu\sigma$

$$B'(e_c) = \alpha[\kappa_\mu n D'(ne_c) + (1-\kappa_\mu) m D'(E)] + (1-\alpha)[\kappa_\varepsilon n D'(ne_c) + (1-\kappa_\varepsilon) m D'(E)], \tag{26a}$$

$$B'(e_f) = \alpha[\kappa_\mu D'(ne_f) + (1-\kappa_\mu) D'(E)] + (1-\alpha)[\kappa_\varepsilon n D'(ne_f) + (1-\kappa_\varepsilon) D'(E)], \tag{26b}$$

[19] In the grey-shaded areas northeast of the green lines of Figure 7 - 9 it holds $S_{\varepsilon\mu\sigma}(m^* = n; \cdot) > 0$.

where $E := me_c + (n-m)e_f$, we infer that an increase in κ_ε ceteris paribus enhances the fringe countries' internalization of the damage by the factor $(1-\alpha)(n-1)D'$ whereas it increases the coalition countries internalization of the damage by the factor $(1-\alpha)(n-m)D'$. Thus, the fringe country's free-riding incentives are declining in κ_ε. Next, we increase κ_ε from $\kappa_\varepsilon = 0$ to some $\kappa_\varepsilon > 0$ in an economy without membership moralists ($\alpha = 0$) and in an economy with membership moralists ($\alpha > 0$). The Figures 11 and 12 illustrate for the numerical values $\alpha = 0.5$, $\beta = 200$, $\delta = 1$, $\kappa_\varepsilon = 0.5$, $\kappa_\mu = 0.3$ and $n = 100$ how the equilibrium emissions and welfare levels depend on the coalition size. The difference[20] $\hat{E}^f_{\varepsilon\mu\sigma} - \hat{E}^c_{\varepsilon\mu\sigma}$ captures the free-riding incentives of fringe countries. The left panels of the Figures 11 and 12 show that increasing κ_ε reduces the free-riding incentives and this reduction is the stronger the larger is the coalition.

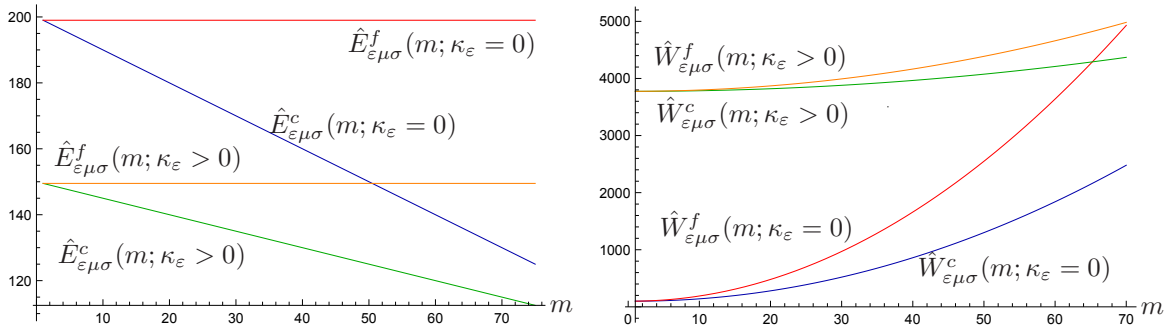

Figure 11: Increases of κ_ε in an economy without membership moralists

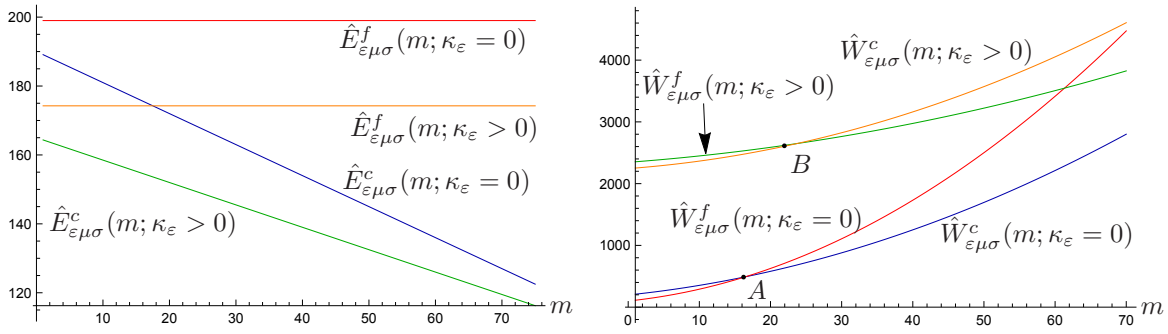

Figure 12: Increases of κ_ε in an economy with membership moralists

[20] In Lemma 1 of the Appendix it is shown that $\hat{E}^f > \hat{E}^c$ holds in all games except in Game ε and Game μ.

In an economy without membership moralists ($\alpha = 0$), in which the size of the stable coalition is $m = 3$, increasing κ_ε reduces the free-riding incentives only slightly, such that the new stable coalition size remains at three (see right panel of Figure 11). In contrast, in an economy with membership moralists ($\alpha > 0$) and without emissions Kantians ($\kappa_\varepsilon = 0$) the stable coalition is already larger (see point A in the right panel of Figure 12). Increasing κ_ε in that economy reduces the free-riding incentives of fringe countries outside the formerly stable coalition very strongly such that the size of the new stable coalition increases. Such an increase is documented in the right panel of Figure 12 by the observation that the intersection point B lies to the right of initial intersection point A. To sum up, the larger the stable coalition in an economy the stronger is the reduction of free-riding incentives caused by an increase in κ_ε, and the greater is the increase in the size of the stable coalition.

3 Concluding remarks

Our motivation to investigate the impact of morally behaving countries on international environmental cooperation is the empirical observation that a growing number of individuals increase their contributions to limit the (pending) global climate damage. They deliberately reduce their own carbon footprint but also develop a moral preference for, or pressure on, their home government to play an active part in the process of international cooperation. We consider such pro-climate behavior to be guided by the moral principle "to do the right thing" even at the cost of some loss of material well-being. If sufficiently many individuals exhibit some degree of that kind of morality, they elect governments committed to their electorate's moral attitude. We expect such moral governments to perform better than 'non-moral' governments with regard to the formation of deep and broad climate coalitions.

On the whole, that expectation proved to be correct. In the encompassing IEA Game $\varepsilon\mu\sigma$ increasing degrees of morality and an increasing weight on the membership moralist's welfare yield larger stable coalitions. Without repeating details, it is worth emphasizing that for increasing the size of stable coalitions the (high) degree of membership morality is more important than the degree of emissions morality. Emissions moralists reduce emissions below the level chosen by materialists, but they are indifferent with regard to coalition formation. It is therefore also interesting to note that combined with a given positive degree of membership morality an increasing degree of emissions morality does increase the size of the stable coalition in the encompassing Game $\varepsilon\mu\sigma$.

In this paper, we chose a parsimonious theoretical framework for the benefit of clear-cut results and their drivers. In the long list of desiderata for future research are items

such as the consideration of numerous consumption goods with different carbon emissions intensities; heterogeneous countries and a more explicit and realistic link between moral consumers and moral governments. On the empirical side, a formidable task beyond the scope of the present paper is to identify the kind of morally motivated behavior we discussed and to suggest suitable operational measurement procedures of our moral parameters.

References

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 (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.

Ansink, E., Weikard, H.-P. and C. Withagen (2019): International environmental agreements with support, *Journal of Environmental Economics and Management* 97, 241-252.

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

Bayramoglu, B., Jacques, J.-F. and M. Finus (2018): Climate agreements in a mitigation-adaptation game, *Journal of Public Economics* 165, 101-113.

Bernauer, T., Gampfer, R., Meng, T. and Y.-S. Su (2016): Could more civil society involvement increase public support for climate policy-making? Evidence from a survey experiment in China, *Global Environmental Change*, 40, 1-12.

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.

Dasgupta, P., Southerton, A., Ulph. A. and D. Ulph (2016): Consumer behavior with environmental and social externalities: Implications for analysis and policy, *Environmental and Resource Economics* 65, 191-226.

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

De Zeeuw, A. (2018): Dynamic effects on the stability of international environmental agreements, *Journal of Environmental Economics and Management* 55, 163-174.

Diamantoudi, E. and E. Sartzetakis (2016): International environmental agreements under foresight, *Economic Theory* 59, 527-546.

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

Eichner, T. and R. Pethig (2015): Is trade liberalization conducive to the formation of climate coalitions?, *International Tax and Public Finance* 22, 932-955.

Eichner, T. and R. Pethig (2013): Self-enforcing environmental agreements, *Journal of Public Economics* 102, 37-50.

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.

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

Finus, M. and D. Rübbelke (2012): Public good provision with ancillary effects: The case of climate agreements, *Environmental and Resource Economics* 56, 211-226.

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

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.

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

Jakob, M., Kübler, D., Steckel, J.C. and R. van Veldhuizen (2017): Clean up your own mess: An experimental study of moral responsibility and efficiency, *Journal of Public Economics* 155, 138-246.

Kant, I. (1785): Grundlegung zur Metaphysik der Sitten. [In English: Groundwork of the Metaphysics of Morals. 1964. New York: Harper Torch books.]

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.

McEvoy, D.M. and M. McGinty (2018): Negotiating a uniform emissions tax in international environmental agreements, *Journal of Environmental Economics and Management* 90, 217-231.

Nyborg, K. (2018a): Social norms and the environment, *Annual Review of Resource Economics* 10, 405-423.

Nyborg, K. (2018b): 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.

UNEP (United Nations Environment Programme) (2019): The emissions gap report 2019.

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 Long, N. (2016): The impacts of the other-regarding preferences and ethical choice on environmental outcomes: A review of the literature, Scientific Series 2016s-10, CIRANO, Montreal.

Van der Pol, T., Weikard, H.-P. and E. van Ireland (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.

Appendix

Proof of Proposition 5:

(i) Using the notation $e_i = e_c$ for all $i \in C(\mathbf{s})$ and $e_i = e_f$ for all $i \in F(\mathbf{s})$, in Game $\mu\sigma$ the first-order conditions are

$$B'(e_c) - \kappa_\mu n D'(ne_c) - (1-\kappa_\mu) m D'(E) = 0, \tag{A1a}$$

$$B'(e_f) - \kappa_\mu D'(ne_f) - (1-\kappa_\mu) D'(E) = 0, \tag{A1b}$$

where $E = me_c + (n-m)e_f$. Denoting the solution of (A1a) and (A1b) by $e_c^* = \hat{E}^c_{\mu\sigma}(m;\kappa_\mu)$ and $e_f^* = \hat{E}^f_{\mu\sigma}(m;\kappa_\mu)$, the welfare levels are

$$\hat{W}^c_{\mu\sigma}(m;\kappa_\mu) = B(e_c^*) - \kappa_\mu D(ne_c^*) - (1-\kappa_\mu) D(E^*), \tag{A2a}$$

$$\hat{W}^f_{\mu\sigma}(m;\kappa_\mu) = B(e_f^*) - \kappa_\mu D(ne_f^*) - (1-\kappa_\mu) D(E^*). \tag{A2b}$$

In view of (A2a) and (A2b), $B'' < 0$, $e_c^* < e_f^*$ (see Lemma 1 below) for $\kappa_\mu = 0$ we obtain

$$\hat{W}^f_{\mu\sigma}(m;0) > \hat{W}^c_{\mu\sigma}(m;0). \tag{A3}$$

For $\kappa_\mu = 1$ we get $e_c^* = e^{\mathrm{SO}}$, $e_f^* = e^{\mathrm{BAU}}$ and hence

$$\hat{W}^f_{\mu\sigma}(m;1) = w^{\mathrm{BAU}} < \hat{W}^c_{\mu\sigma}(m;1) = w^{\mathrm{SO}}. \tag{A4}$$

Under the minor restriction that sign $\left(\frac{\partial^2 \hat{W}^c_{\mu\sigma}(m;\kappa_\mu)}{\partial \kappa_\mu^2}\right)$ is the same for all $\kappa_\mu \in [0,1]$ and that sign $\left(\frac{\partial^2 \hat{W}^f_{\mu\sigma}(m;\kappa_\mu)}{\partial \kappa_\mu^2}\right)$ is the same for all $\kappa_\mu \in [0,1]$ it follows (i).

(ii) For the parametric functions (2b) we solve (A1a) and (A1b) to get

$$e_c^* = e^{\mathrm{SO}} + (n-m)(1-\alpha\kappa_\mu)\delta, \tag{A5a}$$

$$e_f^* = e^{\mathrm{SO}} + (n-1)\delta, \tag{A5b}$$

where $e^{\mathrm{SO}} = \beta - n\delta$. Inserting (A5a) and (A5b) into the welfare functions yields

$$\hat{W}^c_{\mu\sigma}(m;\kappa_\mu) = w^{\mathrm{SP}} + \frac{\delta^2(n-m)(1-\kappa_\mu)}{2}\left[n(1+\kappa_\mu) - m(1-\kappa_\mu) - 2\right], \tag{A6a}$$

$$\hat{W}^f_{\mu\sigma}(m;\kappa_\mu) = w^{\mathrm{SP}} - \frac{\delta^2}{2}\left[(n-1)^2 - 2m^2(1-\kappa_\mu)^2 - 2m(1-\kappa_\mu)(n\kappa_\mu - 1)\right], \tag{A6b}$$

where $w^{\mathrm{SO}} = \frac{1}{2}(\beta - n\delta)^2$. Making use of (A6a) and (A6b) in the stability function and rearranging terms we get

$$S_{\mu\sigma}(m;\kappa_\mu) = \frac{\delta^2}{2}\left[(n^2-2n-2)\kappa_\mu^2 - 3(1-2\kappa_\mu) - (m^2-4m)(1-\kappa_\mu)^2\right]. \tag{A7}$$

Solving $S_{\mu\sigma}(m;\kappa_\mu) = 0$ with respect to m yields

$$m = 2 + \frac{\sqrt{(1-\kappa_\mu)^2 + (n-1)^2\kappa_\mu^2}}{1-\kappa_\mu} =: \mathring{M}(\kappa_\mu, n) \tag{A8}$$

Proposition 5(ii) follows from $\mathring{M}(0,n) = 3$, $\mathring{M}_{\kappa_\mu} = \frac{(n-1)^2}{(1-\kappa_\mu)^2\sqrt{(1-\kappa_\mu^2)+(n-1)^2\kappa_\mu^2}} > 0$ and $\mathring{M}(\tilde{\kappa}_\mu, n) = n$ for $\tilde{\kappa}_\mu = \frac{(n-3)-\sqrt{(n-3)(n-1)}}{2}$. ■

Proof of Proposition 6:

(i) Assuming that the functions B and D take the parametric functional forms (2b), solving (26a) and (26b) yields

$$e_c^* = e^{\mathrm{SP}} + (n-m)[1-(1-\alpha)\kappa_\varepsilon - \alpha\kappa_\mu]\delta, \tag{A9a}$$

$$e_f^* = e^{\mathrm{SP}} + (n-1)[1-(1-\alpha)\kappa_\varepsilon]\delta, \tag{A9b}$$

where $e^{\text{SP}} = \beta - n\delta$. Inserting the emission levels e_c^* and e_f^* into the welfare functions which in turn are inserted into the stability function we get

$$\begin{aligned} S_{\varepsilon\mu\sigma}(m;\alpha,\kappa_\varepsilon,\kappa_\mu) &= -\frac{\delta^2}{2}\left[3+3(1-\alpha)^2\kappa_\varepsilon^2-6\alpha\kappa_\mu+(2+2n-n^2)\alpha^2\kappa_\mu^2\right. \\ &\quad \left.+(m^2-4m)\left((1-\alpha)\kappa_\varepsilon+\alpha\kappa_\mu-1\right)^2-6(1-\alpha)\kappa_\varepsilon(1-\alpha\kappa_\mu)\right] \quad \text{(A10)} \end{aligned}$$

Solving $S_{\varepsilon\mu\sigma}(m;\alpha,\kappa_\varepsilon,\kappa_\mu)=0$ with respect to m yields

$$m = \frac{2\psi+[1-(1-\alpha)\kappa_\varepsilon-\alpha\kappa_\mu]\sqrt{\psi}}{[1-(1-\alpha)\kappa_\varepsilon-\alpha\kappa_\mu]^2} =: M(\alpha,\kappa_\varepsilon,\kappa_\mu,n), \quad \text{(A11)}$$

where $\psi := [(1-\alpha)\kappa_\varepsilon+\alpha\kappa_\mu]^2+1-2(1-\alpha)\ \kappa_\varepsilon-2\alpha\kappa_\mu$. Differentiation of (A11) leads to

$$M_{\kappa_\varepsilon} = \frac{(n-1)^2(1-\alpha)\alpha^2\kappa_\mu^2}{[1-(1-\alpha)\kappa_\varepsilon-\alpha\kappa_\mu]^2\sqrt{\psi}} > 0, \quad \text{(A12a)}$$

$$M_{\kappa_\mu} = \frac{(n-1)^2\alpha^2\kappa_\mu[1-(1-\alpha)\kappa_\varepsilon]}{[1-(1-\alpha)\kappa_\varepsilon-\alpha\kappa_\mu]^2\sqrt{\psi}} > 0 \quad \text{(A12b)}$$

$$M_\alpha = \frac{(n-1)^2\alpha(1-\kappa_\varepsilon)\kappa_\mu^2}{[1-(1-\alpha)\kappa_\varepsilon-\alpha\kappa_\mu]^2\sqrt{\psi}} > 0, \quad \text{(A12c)}$$

$$\frac{\partial M\left(\frac{1}{2},\kappa,\kappa,n\right)}{\partial\kappa} = \frac{(n-1)^2\kappa}{2(1-\kappa)^2\sqrt{4(1-\kappa)^2+(n-1)^2\kappa^2}} > 0. \quad \text{(A12d)}$$

(ii) Consider the grand coalition which is stable if and only if

$$\begin{aligned} S_{\varepsilon\mu\sigma}(n;\alpha,\kappa_\varepsilon,\kappa_\mu) &= -\frac{\delta^2}{2}\left[(n^2-4n+3)[(1-(1-\alpha)\kappa_\varepsilon)^2-2\alpha\kappa_\mu(1-(1-\alpha)\kappa_\varepsilon)\right. \\ &\quad \left.-2(n-1)\alpha^2\kappa_\mu^2\right] \geq 0. \quad \text{(A13)} \end{aligned}$$

Differentiation of (A13) leads to

$$\frac{\partial S_{\varepsilon\mu\sigma}}{\partial\kappa_\varepsilon} = \delta^2(n-3)(n-1)(1-\alpha)[1-(1-\alpha)\kappa_\varepsilon-\alpha\kappa_\mu] > 0, \quad \text{(A14a)}$$

$$\frac{\partial S_{\varepsilon\mu\sigma}}{\partial\kappa_\mu} = \delta^2(n-3)(n-1)[1-(1-\alpha)\kappa_\varepsilon+2\alpha\kappa_\mu] > 0, \quad \text{(A14b)}$$

$$\frac{\partial S_{\varepsilon\mu\sigma}\left(n;\frac{1}{2},\kappa,\kappa\right)}{\partial\kappa} = \frac{\delta^2}{2}\left[4(n-3)(n-1)-\kappa\left(3n^2-14n+11\right)\right] > 0. \quad \text{(A14c)}$$

The sign in (A14c) follows from $G_\kappa = -(3n^2-14n+11) < 0$ and $G(n;1) = (n-1)^2$, where $G(n;\kappa) := 4(n-3)(n-1)-\kappa\,(3n^2-14n+11)$.

Finally, $S_{\varepsilon\mu\sigma}(n;0,\kappa_\varepsilon,\kappa_\mu) = -\frac{\delta^2}{2}(n-3)(n-1)(1-\kappa_\varepsilon)^2 < 0$,

$$\begin{aligned} &S_{\varepsilon\mu\sigma}(n;\tilde{\alpha},\kappa_\varepsilon,\kappa_\mu) = 0 \\ &\qquad \Longleftrightarrow \quad \tilde{\alpha} = \frac{(n-3)[\kappa_\varepsilon^2-\kappa_\varepsilon(1+\kappa_\mu)+\kappa_\mu]-(1-\kappa_\varepsilon)\kappa_\mu\sqrt{(n-3)(n-1)}}{(n-3)(\kappa_\varepsilon^2-2\kappa_\varepsilon\kappa_\mu)-2\kappa_\mu^2} \quad \text{(A15)} \end{aligned}$$

and $\left.\frac{\partial S_{\varepsilon\mu\sigma}}{\partial\alpha}\right|_{\alpha=\tilde{\alpha}} = (n-1)^2\delta^2(1-\kappa_\varepsilon)\kappa_\mu\sqrt{(n-3)(n-1)} > 0$ proves Proposition 6(iic). ■

Lemma 1. *For any given $m \in [2, n-1]$, $\alpha < 1$, $\kappa_\varepsilon < 1$ and $\kappa_\mu < 1$ it holds*

$$e_c^* < e_f^*. \tag{A16}$$

Proof: The proof is by contradiction. From (26a) and (26b) we get

$$\frac{B'(e_c) - \alpha\kappa_\mu n D'(ne_c) - (1-\alpha)\kappa_\varepsilon n D'(ne_c)}{m} = B'(e_f) - \alpha\kappa_\mu D'(ne_f) - (1-\alpha)\kappa_\varepsilon n D'(ne_f). \tag{A17}$$

Define $\tilde{B}(e_q) = B(e_q) - \alpha\kappa_\mu D(ne_q) - (1-\alpha)\kappa_\varepsilon D(ne_q)$ for $q = c, f$. Accounting for $\tilde{B}' = B' - \alpha n\kappa_\mu D'(ne_q) - (1-\alpha)n\kappa_\varepsilon D'(ne_q)$ in (A17) we obtain

$$\frac{\tilde{B}'(e_c)}{m} = \tilde{B}'(e_f) + \alpha(n-1)\kappa_\mu D'(ne_f). \tag{A18}$$

Suppose $e_c > e_f$. Then $\tilde{B}'' < 0$ implies $\tilde{B}'(e_c) < \tilde{B}'(e_f)$ and $\frac{\tilde{B}'(e_c)}{m} < \tilde{B}'(e_f) + \alpha(n-1)\kappa_\mu D'(ne_f)$, which contradicts (A18). ■