ECONSTOR
Make Your Publications Visible.

A Service of

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Schopf, Mark

Article — Published Version

Self-Enforcing International Environmental Agreements and Altruistic Preferences

Environmental and Resource Economics

Provided in Cooperation with:
Springer Nature

Suggested Citation: Schopf, Mark (2024) : Self-Enforcing International Environmental Agreements and Altruistic Preferences, Environmental and Resource Economics, ISSN 1573-1502, Springer Netherlands, Dordrecht, Vol. 87, Iss. 9, pp. 2309-2359, https://doi.org/10.1007/s10640-024-00885-8

This Version is available at:
https://hdl.handle.net/10419/315261

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

http://creativecommons.org/licenses/by/4.0/

WWW.ECONSTOR.EU

Mitglied der Leibniz-Gemeinschaft

Environmental and Resource Economics (2024) 87:2309–2359
https://doi.org/10.1007/s10640-024-00885-8

Self-Enforcing International Environmental Agreements and Altruistic Preferences

Mark Schopf[1]

Accepted: 10 May 2024 / Published online: 25 June 2024
© The Author(s) 2024

Abstract
This paper analyses the effects of altruism on the formation of climate coalitions in the standard two-stage game of self-enforcing international environmental agreements with identical countries. Altruism implies that each country values, to some extent, every other country's welfare when deciding on its coalition membership and emissions policy. In the Nash [Stackelberg] game, the fringe [coalition] countries exploit the altruism of the coalition [fringe] countries so that altruism decreases [increases] the coalition size. In any case, global emissions and global welfare are close to the non-cooperative values. However, altruism narrows the gap between the individually optimal emissions and the socially optimal emissions, so altruism increases global welfare. The effects of altruism on the formation of climate coalitions crucially depends on its modelling: If altruism affects the membership decision but not the policy decision, or if each coalition country is more altruistic toward other coalition countries than toward fringe countries, altruism can stabilise large coalitions up to the grand coalition. Finally, altruism can stabilise small coalitions but destabilises large coalitions with asymmetric countries.

Keywords Climate coalition · Climate policy · Moral behaviour · Social norms

JEL Classification C72 · D64 · Q54 · Q58

1 Introduction

The Paris Agreement, negotiated by 196 parties at the 2015 United Nations Climate Change Conference, aims to limit global warming to well below 2 °C compared to pre-industrial levels (UN 2015). Although there is thus broad consensus on the international goal of climate policy, a continuation of current policies would result in global warming of about 3 °C above pre-industrial levels (UN 2023).[1] Consequently, the Paris Agreement with its nationally determined contributions does not constitute an effective international

[1] If current policies continue through 2030 and the implied carbon price in 2030 increases with the global growth rate through 2100, there is a 50% [4%] median chance of limiting global warming to 2.7 °C

✉ Mark Schopf
mark.schopf@fernuni-hagen.de

[1] Department of Economics, University of Hagen, Universitätsstr. 41, 58097 Hagen, Germany

environmental agreement in terms of the international political narrative. On the other hand, some world regions have introduced rather high carbon prices despite facing negative social costs of carbon (see Table 1). Although these carbon prices are still well below the global social cost of carbon (418/tCO_2 from Ricke et al. 2018), this behaviour can hardly be explained with perfect selfishness.[2]

Instead, it may reflect the important effects of altruistic values on environmental behaviour found in the psychological literature (see, e.g., Dietz et al. 2005; Steg 2016; Lades et al. 2021).[3] In particular, there is evidence that climate policy negotiators have social preferences regarding burden-sharing rules: If climate policy negotiators from rich countries were perfectly selfish, they would support the grandfathering rule, i.e. equal percentage reduction of emissions, and oppose the egalitarian rule, i.e. equal per capita emissions. However, gross domestic product per capita is not positively [negatively] correlated with support for the grandfathering [egalitarian] rule (Lange et al. 2007; Meulemann and Ziegler 2015). Furthermore, climate policy negotiators from industrialized countries state that the egalitarian rule should be the most important burden-sharing rule in international climate agreements (Kesternich et al. 2021). These results suggest that climate policy negotiators are not driven by perfect selfishness.[4]

This paper analyses the formation of climate coalitions with altruistic preferences. In particular, each country values, to some extent, every other country's welfare when deciding on its coalition membership at the first stage of the game and emissions policy at the second stage of the game. In order to be able to compare our results with the standard literature (Carraro and Siniscalco 1993; Barrett 1994), we apply the canonical model of self-enforcing international environmental agreements with identical countries, concave utility from own emissions and convex costs from global emissions. Without altruistic preferences, the standard (Nash or Stackelberg) game with linear-quadratic emissions benefits and linear marginal emissions damages predicts either small or ineffective climate coalitions.[5] Beyond the linear-quadratic case, Barrett (2013) and Nkuiya et al. (2015) show that climate thresholds can stabilise the grand coalition, Nkuiya (2020) finds that

Footnote 1 (continued)

[2 °C] above pre-industrial levels (UN 2023, Chapter 4.5, Appendix C). However, the very high emissions scenario of the IPCC (2021), which best reflects the cumulative emissions from 2005 to 2020 (Schwalm et al. 2020), even predicts a temperature increase of 4.4 °C above pre-industrial levels (median; fifth to ninety-fifth percentile: 3.3–5.7 °C) by the end of the century (IPCC 2021, Chapter 4.3).

[2] With strategic climate policy, terms-of-trade effects could explain part of the gap between regional carbon prices and regional social costs of carbon (Markusen 1975; Hoel 1996). Furthermore, the presence of an international carbon market with endogenous permit choice (Helm 2003; Holtsmark and Weitzman 2020) incentivises regions with low abatement costs to mitigate emissions and sell permits, which can reduce global emissions and raise global welfare. This has been shown for an exogenous carbon market with an endogenous climate coalition (Altamirano-Cabrera and Finus 2006; Lessmann et al. 2014), an endogenous carbon market without climate coalition (Carbone et al. 2009; Holtsmark and Midttømme 2021) and an endogenous carbon market with an endogenous climate coalition (Yu and Wu 2022).

[3] Further examples include Kotchen and Moore (2007) and Ziegler (2020) [Engler et al. (2022) and Andre et al. (2024)], who find that altruistic values are significantly positively correlated with participation in green-electricity programs [pro-climate donations].

[4] Hjerpe et al. (2011) argue that the ability-to-pay rule, in terms of gross domestic product per capita, "has the greatest potential to serve as a basis for agreement in negotiations on allocating mitigation commitments" because it is supported by many and opposed by few climate negotiators. Furthermore, it is not opposed by any climate negotiator from industrialized countries.

[5] In the Nash game with linear-quadratic emissions benefits and quadratic emissions damages, Finus (2001, p. 232) finds that climate coalitions consist of no more than two countries. In the Stackelberg game with linear-quadratic emissions benefits and quadratic emissions damages, Finus (2001, p. 232) finds that

climate coalitions can be large and effective with isoelastic emissions benefits in the Stackelberg game, and Eckert and Nkuiya (2022) show that convex marginal emissions damages can stabilise large coalitions up to the grand coalition in the Nash game. With general functional forms, the second stage of the Nash game and the Stackelberg game have been analysed in detail by Bayramoglu et al. (2018) and Finus et al. (2021a, 2021b), respectively. Finally, Finus et al. (2023) show that the stable coalition is always weakly larger in the Stackelberg game than in the Nash game.[6]

We distinguish between the coalition countries taking the fringe countries' emissions as given (Nash game) and taking the reaction of the fringe countries' emissions into account (Stackelberg game) when choosing their own emissions. In both cases, altruism reduces each fringe country's emissions and raises global material welfare, i.e. global welfare in the absence of altruistic preferences. Furthermore, we get the typical results that global emissions decrease and each fringe country's emissions and material welfare increase with the coalition size in the Nash game and above a critical coalition size in the Stackelberg game. By contrast, the effect of altruism on the equilibrium coalition size depends crucially on the game structure.

In the Nash game with linear-quadratic emissions benefits and quadratic emissions damages, altruism weakly reduces the coalition size, and climate coalitions consist of no more than two countries. The direct effect of altruism, namely smaller global emissions and larger global material welfare for a larger coalition size, makes it worthwhile for all other countries if some country joins the coalition. However, the indirect effect of altruism, namely smaller global emissions and larger global material welfare for a given coalition size, makes it less costly for all other countries if some country does not join the coalition. This indirect effect outweighs the direct effect for small coalition sizes and explains the small climate coalition in equilibrium.

In the Stackelberg game with linear-quadratic emissions benefits and quadratic emissions damages, altruism weakly raises the coalition size, and climate coalitions can consist of up to six countries. In this case, the coalition countries take advantage of the fringe countries' altruism by becoming less ambitious in the fight against climate change, expecting the fringe countries to react by reducing their emissions more than they would without altruism. However, the coalition countries are not much more ambitious in the Stackelberg equilibrium than in the business-as-usual scenario without coalition formation.

These results suggest that altruism cannot stabilise large and effective climate coalitions. However, altruism narrows the gap between the individually optimal emissions and the socially optimal emissions, so altruism increases global welfare. Thus, altruism affecting the membership decision and the policy decision appears to be more of a substitute than a complement for large climate coalitions.

The economic literature has developed and tested several theories for imperfect selfishness. In the case of altruistic preferences (Becker 1974), one can distinguish between pure altruism, i.e. utility from others' utility values (Becker 1981), paternalistic altruism, i.e. utility from others' consumption bundles (Pollak 1988), and impure altruism, i.e. utility or warm glow from giving others (Andreoni 1990). Alger and Weibull (2010) show that pure altruism used in this paper is evolutionary stable, and Andreoni et al. (2010)

Footnote 5 (continued)

climate coalitions are either small or ineffective, and Diamantoudi and Sartzetakis (2006, p. 254) find that climate coalitions consist of no more than four countries when constraining the parameter space to ensure non-negative emissions.

[6] With linear abatement benefits, Karp and Simon (2013) find that a coalition of two or less [three or more] countries is stable with strictly convex [concave] marginal abatement costs in the Nash game.

Table 1 Largest carbon pricing schemes representing 22% of global CO_2 emissions

$/t$CO_2$	EU	GBR	CAN	USA	KOR	ZAF	CHN	ARG	MEX	JPN	KAZ	UKR
Price	73	58	38	28	19	10	10	5	4	2	1	1
SCC	−4	−4	−8	48	−1	3	24	3	12	6	−1	−1

Price: The World Bank (2023), SCC (social cost of carbon): Ricke et al. (2018)

summarize the significant evidence for altruism in economic experiments. Other theories comprise reciprocal fairness (Rabin 1993), inequality aversion (Fehr and Schmidt 1999; Bolton and Ockenfels 2000) and Kantian behaviour (Alger and Weibull 2013; Roemer 2015).

These theories have also been applied in the literature on self-enforcing international environmental agreements. Buchholz et al. (2018) and Nyborg (2018) analyse the effects of reciprocal fairness when countries decide on their membership in the coalition and on their emissions. They find that reciprocal fairness can stabilise the grand coalition, but it can also stabilise an interior coalition that is either weakly larger (Nyborg 2018) or even weakly smaller (Buchholz et al. 2018) than the interior coalition without reciprocal fairness. Lange and Vogt (2003) incorporate inequality aversion à la Bolton and Ockenfels (2000) into the canonical model of self-enforcing international environmental agreements and find that sufficiently large inequality aversion can stabilise the grand coalition. By contrast, Vogt (2016) applies inequality aversion à la Fehr and Schmidt (1999) and finds no stable coalition without transfers in his numerical model with heterogeneous countries. Recently, Eichner and Pethig (2022) and Ulph and Ulph (2023) analysed the effects of Kantian or moral behaviour when countries decide on their membership in the coalition and on their emissions. They find that membership moralism expands the climate coalition, and emissions moralism can expand the climate coalition only in the presence of membership moralism.

Closest to our paper is van der Pol et al. (2012), who analyse the effects of altruism affecting the membership decision but not the policy decision. They find that this kind of partial altruism expands the climate coalition. We extend their model into different directions. First, we consider altruism on both stages of the game. Second, we analyse not only the Nash game but also the Stackelberg game. Third, while they solve their model numerically with heterogeneous countries, we solve our model analytically with homogeneous countries. Forth, we replicate their results analytically to discuss the differences from our results.[7]

Finally, we perform two model extensions. First, we analyse the effects of community altruism, that is, we distinguish between in-group altruism and out-group altruism. In this case, altruism can stabilise large coalitions up to the grand coalition. Second, we analyse the effects of altruism with linear climate damages and asymmetric countries. In this case, altruism can stabilise small coalitions but destabilises large coalitions.

[7] Daube (2019) and Goussebaïle et al. (2023) analyse the effects of altruism on climate policy with multiple countries. Daube (2019) shows that altruistic preferences lead to a partial internalization of the climate externality in the non-cooperative solution, and to a full internalization of the climate externality in the cooperative solution if and only if the altruistic preferences for all countries coincide. Goussebaïle et al. (2023) analyse the effects of altruistic foreign aid on climate change mitigation and find that paying transfers before abating emissions incentivises developing countries to choose efficient climate change mitigation and leads to the social optimum if altruistic preferences are sufficiently large. However, both papers abstract from coalition formation.

The remainder of the paper is organized as follows: Sect. 2 introduces the model, and characterises the social optimum and the business-as-usual scenario. Section 3 analyses the effects of altruism on the Nash game of coalition formation. This section also includes a comparison with the model of van der Pol et al. (2012). Section 4 analyses the effects of altruism on the Stackelberg game of coalition formation with the coalition countries as Stackelberg leaders and the fringe countries as Stackelberg followers. Section 5 discusses which realms of decision making might be influenced by social preferences. Section 6 performs our two model extensions. Section 7 concludes.

2 Model

Consider a model with $n \geq 3$ identical countries.[8] Each county $i \in N$ derives consumption benefits $B(e_i)$ from its emissions e_i, where $B(0) \geq 0$, $B' > 0$ and $B'' < 0$, and faces climate damages $D(e)$ from global emissions $e := \sum_{i \in N} e_i$, where $D(0) = 0$, $D' > 0$ and $D'' > 0$. Then, each country's material welfare function is $W_i = B(e_i) - D(e)$. Furthermore, each country is altruistic such that it values its own material welfare by 1 and every other country's material welfare by $\alpha \in [0, 1]$.[9] Thus, the altruism parameter $\alpha = 0$ implies perfectly selfish countries, while $\alpha = 1$ implies perfectly altruistic countries. Then, each country's moral welfare is

$$V_i = W_i + \alpha \sum_{j \in N \setminus i} W_j = (1 - \alpha) W_i + \alpha W, \tag{1}$$

where $W := \sum_{i \in N} W_i$ is global material welfare, and the global moral welfare is

$$V := \sum_{i \in N} V_i = \sum_{i \in N} \left[W_i + \alpha \sum_{j \in N \setminus i} W_j \right] = [1 + \alpha(n-1)] W. \tag{2}$$

Consequently, the socially optimal emissions (SO) are independent of the altruism parameter α, while the individually optimal emissions, i.e. the business-as-usual emissions (BAU), are not (Daube 2019, Results 4 and 5). In particular, the socially optimal values and the individually optimal values coincide for $\alpha = 1$. In Appendix A.1, we prove that global emissions decrease and global material welfare increases with the altruism parameter in the individually optimal solution. Consequently, the relative global emissions $e^{\mathrm{BAU}}/e^{\mathrm{SO}}$ decrease and the relative global material and moral welfare $W^{\mathrm{BAU}}/W^{\mathrm{SO}} = V^{\mathrm{BAU}}/V^{\mathrm{SO}}$ increase with the altruism parameter.

In the further course of the paper we analyse the two-stage game of self-enforcing environmental agreements. At the first stage of the game, countries decide on their membership in the coalition. Thereby, internal [external] stability implies that no country will leave [join] the coalition if this reduces its moral welfare (D'Aspremont et al. 1983). At the second stage of the game, there is a coalition of m countries, and countries decide on their

[8] We assume identical countries for analytical tractability with convex climate damages. In reality, countries benefit differently from own emissions and suffer differently from global emissions. For an analysis with asymmetric countries and linear climate damages, see Sect. 6.2.

[9] Instead, if each country values its own material welfare by 1 and every other country's moral welfare by $\gamma \in [0, 1/n]$, then each country's moral welfare function is $V_i = W_i + \gamma \sum_{j \in N \setminus i} V_j = \tilde{W}_i + \tilde{\alpha} \sum_{j \in N \setminus i} \tilde{W}_j$ with $\tilde{W}_i = W_i/(1 + \gamma)$ and $\tilde{\alpha} = \gamma/[1 - \gamma(n-1)] \in [0, 1]$, and our results do not change. For an analysis with community altruism (greater degree of in-group altruism than out-group altruism), see Sect. 6.1.

emissions. Thereby, each fringe country maximizes its moral welfare (1), and each coalition country $i \in M$ maximizes the sum of the coalition countries' moral welfare

$$\sum_{i\in M} V_i = \sum_{i\in M}\left[W_i + \alpha \sum_{j\in N\setminus i} W_j\right] = (1-\alpha)\sum_{i\in M} W_i + \alpha m W. \tag{3}$$

Comparing (1) and (3), each fringe country's policy weights its own material welfare by $1-\alpha$ and global material welfare by α, while each coalition country's policy weights the coalition's material welfare by $1-\alpha$ and global material welfare by αm. In the following we distinguish between two game concepts. In Sect. 3, we analyse the Nash game, and in Sect. 4, we analyse the Stackelberg game with the coalition countries as Stackelberg leaders and the fringe countries as Stackelberg followers. The respective game is then solved by backward induction.

3 Nash Game

At the second stage of the Nash game, each fringe country $i = f$ maximizes its moral welfare (1) over its emissions e_f, taking the other countries' emissions as given, which yields

$$B'(e_f) = [1+\alpha(n-1)]D'(e) \leq nD'(e). \tag{4}$$

Each fringe country equates marginal emissions benefits to its own marginal emissions damages $D'(e)$, plus all other countries' marginal emissions damages weighted by the altruism parameter $\alpha(n-1)D'(e)$.

Furthermore, each coalition country $i = c$ maximizes the sum of the coalition countries' moral welfare (3) over its emissions e_c, taking the other countries' emissions as given, which yields[10]

$$B'(e_c) = \frac{1+\alpha(n-1)}{1+\alpha(m-1)} mD'(e) \leq nD'(e). \tag{5}$$

For $\alpha = 0$, each coalition country equates marginal emissions benefits to the coalition countries' marginal emissions damages $mD'(e)$. For $\alpha > 0$, altruism implies that each coalition country accounts for all other countries' marginal emissions damages via $1+\alpha(n-1)$, but it also implies that all other coalition countries account for each coalition country's marginal emissions benefits via $1+\alpha(m-1)$. Note that $B'(e_f) = B'(e_c) = nD'(e)$ for $\alpha = 1$, so the Nash equilibrium and the social optimum then coincide. In the following we focus on $\alpha \in [0,1)$.

Differentiating (4) and (5) yields the slopes of the aggregate reaction functions

$$R'_f := \frac{\mathrm{d}(n-m)e_f}{\mathrm{d}me_c} = -\frac{(n-m)[1+\alpha(n-1)]D''(e)}{(n-m)[1+\alpha(n-1)]D''(e) - B''(e_f)} \in (-1,0), \tag{6}$$

$$R'_c := \frac{\mathrm{d}me_c}{\mathrm{d}(n-m)e_f} = -\frac{m^2[1+\alpha(n-1)]D''(e)}{m^2[1+\alpha(n-1)]D''(e) - [1+\alpha(m-1)]B''(e_c)} \in (-1,0). \tag{7}$$

[10] The second-order conditions are fulfilled.

Consequently, emissions are strategic substitutes, and the slopes of the aggregate reaction functions ceteris paribus increase in absolute terms with the altruism parameter. Intuitively, altruism implies that countries react more sensitive to other countries' emissions changes. Furthermore, we infer

$$\frac{B'(e_c)}{B'(e_f)} = \frac{m}{1+\alpha(m-1)} \in (1, m]. \tag{8}$$

Consequently, each fringe country's emissions are greater than each coalition country's emissions. In Appendix A.2.1, we prove[11]

Proposition 1 (Comparison of Nash equilibrium and BAU)

- $e_c < e_i^{\mathrm{BAU}} < e_f$ and $e < e^{\mathrm{BAU}}$,
- $V_f > V_c$,
- $W_f > W_c, W_i^{\mathrm{BAU}}$.

(8) implies that the coalition countries are ceteris paribus more ambitious in the fight against climate change than at BAU. This results in smaller coalition country's emissions and global emissions, which raises the free-rider incentives and leads to greater fringe country's emissions. Each fringe country's emissions being greater than each coalition country's emissions implies $V_f > V_c$ and $W_f > W_c$. Finally, global emissions being smaller and each fringe country's emissions being greater than at BAU implies $W_f > W_i^{\mathrm{BAU}}$ and, thus, $V_f > V_i^{\mathrm{BAU}}$ if $W_c \geq W_i^{\mathrm{BAU}}$ or if α is sufficiently small.

To prepare the analysis of the first stage of the Nash game, we prove in Appendix A.2.2[12]

Lemma 1 (Effects of coalition size and altruism on emissions and welfare)

- $\frac{\mathrm{d}e_f}{\mathrm{d}m} > 0, \frac{\mathrm{d}e}{\mathrm{d}m} < 0$ and $\frac{\mathrm{d}W_f}{\mathrm{d}m} > 0$,
- $\frac{\mathrm{d}e_f}{\mathrm{d}\alpha} < 0, \frac{\mathrm{d}e}{\mathrm{d}\alpha} < 0$ and $\frac{\mathrm{d}W}{\mathrm{d}\alpha} > 0$.

From the first bullet of the lemma, we get the typical results that each fringe country's emissions increase but global emissions decrease with the coalition size, so free-rider incentives tend to increase as the coalition gets larger. The resulting higher consumption benefits and lower climate damages imply that each fringe country's material welfare increases with the coalition size and, thus, that V_f increases with the coalition size if W_c increases with the coalition size or if α is sufficiently small.

The second bullet of the lemma reveals that each fringe country's emissions and global emissions decrease with the altruism parameter and that global material welfare increases with the altruism parameter. Consequently, the relative global emissions e/e^{SO} decrease and the relative global material and moral welfare $W/W^{\mathrm{SO}} = V/V^{\mathrm{SO}}$ increase with the altruism parameter. Finally, we prove in Appendix A.2.2 that the slope of the fringe countries' aggregate reaction function increases in absolute terms with the altruism parameter for $D''' \leq 0$ and $B''' = 0$. This slope corresponds to the leakage rate to the fringe countries

[11] For $\alpha = 0$, the results of Proposition 1 have been proven by Bayramoglu et al. (2018, p. 110) for the abatement game.

[12] For $\alpha = 0$, the results of Lemma 1 have been proven by Bayramoglu et al. (2018, p. 110) for the abatement game.

$|R_f'|$, and the higher this leakage rate, the greater ceteris paribus the free-rider incentives in the Nash game (Carraro and Siniscalco 1993).

Now we turn to the first stage of the Nash game. First note that $V_f(m) > V_c(m)$ from Proposition 1 implies that if a coalition is externally unstable, i.e. $V_c(m+1) \geq V_f(m)$, then the corresponding expansion of the coalition is accompanied by a Pareto improvement, i.e. $V_f(m+1) > V_c(m+1) \geq V_f(m) > V_c(m)$. For the detailed stability analysis, we use the following specification with linear-quadratic emissions benefits and quadratic emissions damages

$$B(e_i) = ae_i - \frac{b}{2}e_i^2, \quad D(e) = \frac{d}{2}e^2, \quad \text{with } a, b, d > 0. \tag{9}$$

We constrain the parameter space to ensure non-negative emissions for all $m \in [2, n]$, which gives an upper bound for d. In particular, we formulate the following

Assumption 1 $d \leq \bar{d} := \dfrac{4b}{(n-1)^2}$.

This assumption is necessary and sufficient for non-negative emissions for all $m \in [2, n]$ with $\alpha = 0$.[13] In Appendix A.2.3, we then prove

Proposition 2 (Stability of coalitions with policy altruism) *Consider the specification* (9) *and suppose altruism affects the membership decision and the policy decision.*

- Either the coalition $m = 2$ is stable or no coalition is stable.
- The coalition $m = 2$ is [not] stable for $\alpha = 0$, $n \geq 12$ and $d \leq \bar{d}$ [$\alpha = 0$, $n < 12$ and $d = \bar{d}$].
- The coalition size weakly decreases with α.

From the first bullet of the proposition, we get the pessimistic result that at best a coalition of two countries is stable. The result from the second bullet of the proposition that $m = 2$ is stable for $\alpha = 0$ when there is a sufficiently large number of countries may seem counterintuitive. In fact, there are two countervailing effects of n on the internal stability condition. On the one hand, the condition for $m = 2$ to be stable becomes stricter as n increases for given parameter values a, b, d and α, since the number of fringe countries increases with n, which raises marginal climate damages, reduces each coalition country's emissions and, thus, the benefits of remaining in the coalition. On the other hand, the condition for $m = 2$ to be stable also becomes stricter as d increases, since each coalition [fringe] country's emissions decrease [increase] with the damage parameter, which reflects Barrett's (1994) paradox of cooperation. These considerations imply that the upper bound for d to ensure non-negative emissions decreases with n, i.e. $\partial\bar{d}/\partial n < 0$, which relaxes the internal stability condition. This indirect effect of n on the internal stability condition outweighs the direct effect for $m = 2$ and $\alpha = 0$, which explains why the coalition $m = 2$

[13] Assumption 1 is also sufficient for non-negative emissions for all $m \in [2, n]$ with $\alpha \geq 0.75$. Furthermore, $d \leq b\alpha^2/\{(1-\alpha)[1+\alpha(n-1)-\sqrt{1+\alpha(n-1)}]^2\}$ is necessary and sufficient for non-negative emissions for all $m \in [2, n]$ with $\alpha \in (0, 0.75)$.

is stable for $\alpha = 0$ and $n \geq 12$. By contrast, the coalition $m = 2$ is not stable for $\alpha = 0$ and $n < 12$ when d is sufficiently large, i.e. when $d = \bar{d}$.[14]

We use a numerical example to demonstrate that there are economies in which $m = 2$ is not stable for $\alpha > 0$ and $n \geq 12$. Figure 1 depicts each coalition country's minimal emissions[15] (left-hand side figure) and the internal stability condition for $m = 2$ (right-hand side figure) dependent on α. In the numerical example, each coalition country's emissions are positive for all $m \in [2, n]$. Furthermore, $m = 2$ becomes unstable for $\alpha \geq 0.334$. Thus, there are economies in which $m = 2$ is not stable for $\alpha > 0$ and $n \geq 12$.

The third bullet of Proposition 2 and Fig. 1 show that altruism does not stabilise larger coalitions, but even destabilises small coalitions. This is in stark contrast to the numerical analysis of van der Pol et al. (2012), who find that the coalition size increases with the altruism parameter and that the grand coalition becomes stable for $\alpha \geq 0.401$ (without community altruism and transfers). The major difference between their model and our model is that we assume altruistic preferences at both stages of the game, while they assume altruistic preferences only at the first stage of the game. At the second stage of the game, they assume that each fringe country maximizes its material welfare, while each coalition country maximizes the sum of the coalition countries' material welfare.[16] This does not alter the qualitative results at the second stage of the game, i.e. Proposition 1 and the first bullet of Lemma 1 also hold for $\alpha = 0$. However, it alters the qualitative effects of altruism on the internal stability condition. In both models, this internal stability condition reads

$$\begin{aligned} V_c(m) - V_f(m-1) = (1-\alpha)W_c(m) + \alpha W(m) \\ - \left[(1-\alpha)W_f(m-1) + \alpha W(m-1)\right] \geq 0. \end{aligned} \tag{10}$$

In van der Pol et al. (2012), where the policy is independent of α, altruism stabilises coalitions if and only if

$$\begin{aligned} \frac{\partial\left[V_c(m) - V_f(m-1)\right]}{\partial\alpha} = \left[(m-1)W_c(m) + (n-m)W_f(m)\right] \\ - \left[(m-1)W_c(m-1) + (n-m)W_f(m-1)\right] > 0. \end{aligned} \tag{11}$$

This direct effect of altruism is positive if and only if the total material welfare of the other countries decreases when a country leaves the coalition. Then, altruism can induce a country to stay in the coalition even though its own material welfare would increase if it left the coalition. In Appendix A.2.4, we prove that the direct effect is positive for $m = 2$ (and for $m \in [2, n]$ with specification (9)), regardless of whether or not altruistic preferences are assumed at the second stage of the game. However, the magnitude of the direct effect differs between the models. Furthermore, in our model, where the policy depends on α, altruism stabilises coalitions if and only if

[14] Note that the coalition $m = 2$ is stable for $\alpha = 0$ and $n < 12$ when d is sufficiently small, i.e. when $d \leq b(2\sqrt{n^2 - 3n + 3} - n + 4)/(3n^2 - 4n - 4)$.

[15] Using $e_c(m(\alpha), \alpha)$ with $m(\alpha) = \arg\min e_c(m, \alpha)$.

[16] A discussion on which realms of decision making might be influenced by social preference can be found in Sect. 5.

$$\frac{\mathrm{d}\left[V_c(m)-V_f(m-1)\right]}{\mathrm{d}\alpha}=\frac{\partial\left[V_c(m)-V_f(m-1)\right]}{\partial\alpha}$$
$$+(1-\alpha)\frac{\mathrm{d}W_c(m)}{\mathrm{d}\alpha}+\alpha\frac{\mathrm{d}W(m)}{\mathrm{d}\alpha}-\left[(1-\alpha)\frac{\mathrm{d}W_f(m-1)}{\mathrm{d}\alpha}+\alpha\frac{\mathrm{d}W(m-1)}{\mathrm{d}\alpha}\right]>0. \tag{12}$$

The second line of (12) represents the indirect effect of altruism. It is positive if and only if the policy effect of altruism on a country's moral welfare is greater inside than outside the coalition. In Appendix A.2.4, we prove that the policy effect inside the coalition is positive, i.e. $(1-\alpha)\frac{\mathrm{d}W_c(m)}{\mathrm{d}\alpha}+\alpha\frac{\mathrm{d}W(m)}{\mathrm{d}\alpha}>0$, but that the policy effect outside the coalition is also positive for $m=2$ (and for $m\in[2,n-2]$ with specification (9)), i.e. $(1-\alpha)\frac{\mathrm{d}W_f(m-1)}{\mathrm{d}\alpha}+\alpha\frac{\mathrm{d}W(m-1)}{\mathrm{d}\alpha}>0$. Proposition 2 reveals that the latter effect is so strong that altruism raises the free-rider incentives. In other words, the policy effect of altruism is more important for small coalitions than for large coalitions, and so important that the negative indirect effect of altruism outweighs the positive direct effect.

In order to check whether the different results of van der Pol et al. (2012) indeed stem from the different assumption concerning altruistic preferences at the second stage of the game, and not from some other minor differences between the models, we analyse the first stage of the game without altruistic preferences at the second stage of the game. In Appendix A.2.5, we then prove

Proposition 3 (Stability of coalitions without policy altruism) *Consider specification* (9) *and suppose altruism affects the membership decision but not the policy decision.*

- Either some unique coalition is stable or no coalition is stable.
- The coalition $m=2$ is [not] stable for $\alpha=0$, $n\geq 12$ and $d\leq\bar{d}$ [$\alpha=0$, $n<12$ and $d=\bar{d}$].
- The coalition size weakly increases with α.
- The grand coalition is [not] stable for $\alpha\geq 4/7$ and $d\leq\bar{d}$ [$\alpha\leq 3/7$ and $n\geq 12$].

The first bullet of the proposition reveals that there is at most one stable coalition size. The second bullet of Proposition 3 mirrors the second bullet of Proposition 2, since the models with and without altruistic preferences at the second stage of the game coincide for $\alpha=0$. The rest of Proposition 3 confirms the numerical result of van der Pol et al. (2012) that considering altruism only at the first stage of the game stabilises coalitions. In particular, the grand coalition becomes stable when the altruism parameter is greater than a critical value, which lies between 3/7 and 4/7 for $n\geq 12$ and $d\leq\bar{d}$.

Furthermore, in Appendix A.2.5 we prove that global material and moral welfare increase with the coalition size without altruistic preferences at the second stage of the game. While altruism affecting the membership decision only is beneficial for global welfare and for the climate ($\frac{\mathrm{d}e}{\mathrm{d}m}<0$ from Lemma 1) because it expands the climate coalition, altruism affecting the membership decision and the policy decision is beneficial for global welfare and for the climate ($\frac{\mathrm{d}W}{\mathrm{d}\alpha}>0$ and $\frac{\mathrm{d}e}{\mathrm{d}\alpha}<0$ from Lemma 1) because it tightens the climate policy. If the same coalition is stable in both models, e.g. for $\alpha\to 0$ such that $m=2$, then global welfare is larger and global emissions are smaller with than without altruistic preferences at the second stage of the game. By contrast, if the grand coalition is stable without altruistic preferences at the second stage of the game, e.g. for $\alpha\geq 4/7$, then global welfare is larger and global emissions are smaller without than with altruistic preferences

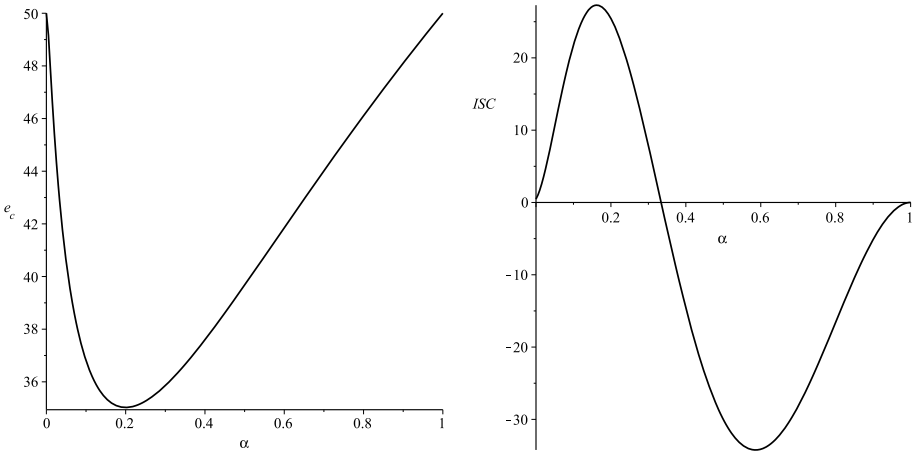

Fig. 1 Each coalition country's minimal emissions (left-hand side figure) and the internal stability condition for $m = 2$ (right-hand side figure) dependent on α with $n = 100, a = 100, b = 1$ and $d = 1/10{,}000$

at the second stage of the game. Figure 2 depicts these relationships for a numerical example.[17] With [without] altruistic preferences at the second stage of the game, $m = 2$ becomes unstable for $\alpha > 0.334$ [$m = n$ becomes stable for $\alpha > 0.5$]. Global emissions are smaller and global material welfare is larger with than without altruistic preferences at the second stage of the game if and only if $\alpha < 0.089$ and $\alpha < 0.135$, respectively. Then, the tighter climate policy outweighs the larger climate coalition, which then comprises no more than 30 and 40 out of 100 countries, respectively. Finally, the figure shows that the welfare difference is relatively small (< 27,000) compared to the welfare difference between social optimum (250,000) and BAU without altruistic preferences (9800).

4 Stackelberg Game

At the second stage of the Stackelberg game, each fringe country $i = f$ maximizes its moral welfare (1) over its emissions e_f, taking the other countries' emissions as given, which yields (4).

Furthermore, each coalition country $i = c$ maximizes the sum of the coalition countries' moral welfare (3) over its emissions e_c, taking the other coalition countries' emissions as given, but taking (4) into account, which yields[18]

$$B'(e_c) = \frac{1+\alpha(n-1)}{1+\alpha(m-1)} mD'(e)\left[1+(1-\alpha)R_f'\right] \leq nD'(e), \tag{13}$$

where $R_f' \in (-1, 0)$ from (6). For $\alpha = 0$, each coalition country equates marginal emissions benefits to the coalition countries' marginal emissions damages $mD'(e)$, corrected for the leakage rate to the fringe countries $|R_f'|$. For $\alpha > 0$, altruism implies that each coalition country accounts for all other countries' marginal emissions damages via $1 + \alpha(n-1)$, but it also implies that all other coalition countries account for each coalition country's marginal emissions benefits via $1 + \alpha(m-1)$. Furthermore, altruism implies that each coalition

[17] Approximating the coalition size by the solution to $V_c(m) - V_f(m-1) = 0$ without altruistic preferences at the second stage of the game.

[18] The second-order conditions are fulfilled if $B''' \geq 0$ and $D''' \leq 0$ (see Appendix A.3.1).

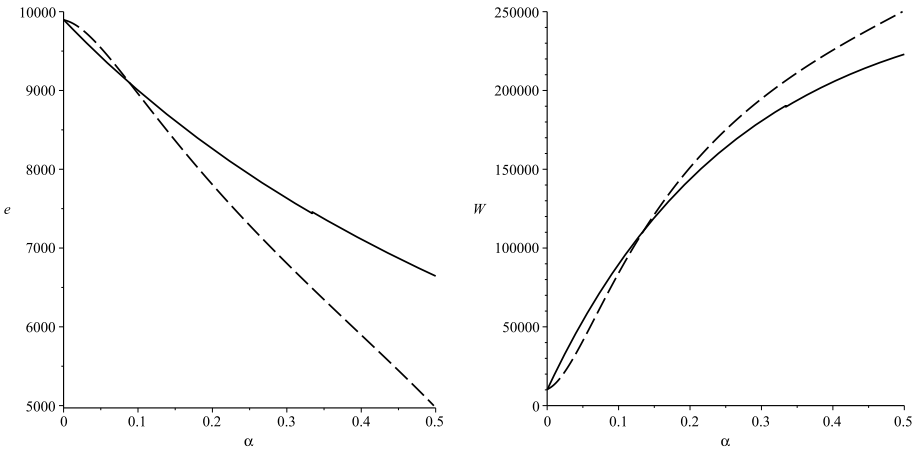

Fig. 2 Global emissions (left-hand side figure) and global material welfare (right-hand side figure) with (solid curves) and without (dashed curves) altruistic preferences at the second stage of the game dependent on α with $n = 100, a = 100, b = 1$ and $d = 1/10{,}000$

country accounts for all fringe countries' marginal emissions benefits, which reduces the influence of the leakage rate to the fringe countries via $1 - \alpha$. Finally, altruism of the fringe countries implies that these countries react more sensitive to other countries' emissions changes, such that the altruism parameter ceteris paribus increases the leakage rate to the fringe countries. Note that $B'(e_f) = B'(e_c) = nD'(e)$ for $\alpha = 1$, so the Stackelberg equilibrium and the social optimum then coincide. In the following we focus on $\alpha \in [0, 1)$.

From (4) and (13), we infer

$$\frac{B'(e_c)}{B'(e_f)} = \frac{m}{1 + \alpha(m-1)}\left[1 + (1-\alpha)R_f'\right] =: \tilde{\theta} \in (0, m). \tag{14}$$

Consequently, each fringe country's emissions are greater [smaller] than each coalition country's emissions for $\tilde{\theta} > [<]1$. Furthermore, $\tilde{\theta} = 1$ implies that the Stackelberg equilibrium and the BAU coincide. In Appendix A.3.2, we prove[19]

Proposition 4 (Comparison of Stackelberg equilibrium and BAU)

- $e_c \gtreqless e_i^{\mathrm{BAU}} \gtreqless e_f$ and $e \gtreqless e^{\mathrm{BAU}}$ for $\tilde{\theta} \lesseqgtr 1$,
- $V_c > V_i^{\mathrm{BAU}} > V_f$ and $V < V^{\mathrm{BAU}}$ for $\tilde{\theta} < 1$, $V_c = V_i^{\mathrm{BAU}} = V_f$ and $V = V^{\mathrm{BAU}}$ for $\tilde{\theta} = 1$, $V_f > V_c > V_i^{\mathrm{BAU}}$ and $V > V^{\mathrm{BAU}}$ for $\tilde{\theta} > 1$,
- $W_c > W_i^{\mathrm{BAU}} > W_f$ and $W < W^{\mathrm{BAU}}$ for $\tilde{\theta} < 1$, $W_c = W_i^{\mathrm{BAU}} = W_f$ and $W = W^{\mathrm{BAU}}$ for $\tilde{\theta} = 1, W_f > W_c, W_i^{\mathrm{BAU}}$ and $W > W^{\mathrm{BAU}}$ for $\tilde{\theta} > 1$.

$\tilde{\theta} > [<]1$ implies that the coalition countries are ceteris paribus more [less] ambitious in the fight against climate change than at BAU. This results in smaller [greater] coalition

[19] For $\alpha = 0$, the results of Proposition 4 have been proven by Finus et al. (2021b, Proposition 2) for the abatement game.

country's emissions and global emissions, which raises [reduces] the free-rider incentives and leads to greater [smaller] fringe country's emissions. The coalition could always choose $\tilde{\theta} = 1$, such that $\tilde{\theta} \neq 1$ implies $V_c > V_i^{\text{BAU}}$. For $\tilde{\theta} > [<]1$ global emissions being smaller [greater] and each fringe country's emissions being greater [smaller] than at BAU implies $V_f \gtreqless V_i^{\text{BAU}} \iff W_f \gtreqless W_i^{\text{BAU}} \iff \tilde{\theta} \gtreqless 1$. Furthermore, for $\tilde{\theta} > [<]1$ global emissions being smaller [greater] than at BAU implies $V \gtreqless V^{\text{BAU}} \iff W \gtreqless W^{\text{BAU}} \iff \tilde{\theta} \gtreqless 1$. Finally, for $\tilde{\theta} > [<]1$ each fringe country's emissions being greater [smaller] than each coalition country's emissions implies $V_f \gtreqless V_c \iff W_f \gtreqless W_c \iff \tilde{\theta} \gtreqless 1$.

The partial derivative of $\tilde{\theta}$ with respect to m is positive, so the coalition countries tend to become more ambitious as the coalition gets larger. Then, the leakage rate to the fringe countries ceteris paribus becomes smaller, which tends to increase $\tilde{\theta}$, see (6). Furthermore, the coalition countries' marginal emissions damages then become greater, which outweighs the greater coalition countries' marginal emissions benefits and increases $\tilde{\theta}$, see (14). In Appendix A.3.3, we prove[20]

Proposition 5 (Relation between coalition size and coalition's ambition) *Suppose* $B''' \geq 0$ *and* $D''' \leq 0$. *Then*, $m \gtreqless \tilde{m} \iff \tilde{\theta} \gtreqless 1$, *where*

$$\tilde{m} := \frac{n[1+\alpha(n-1)]D''(e^{\text{BAU}}) - B''(e_i^{\text{BAU}})}{[1+\alpha(n-1)]D''(e^{\text{BAU}}) - B''(e_i^{\text{BAU}})} \in (1, n) \tag{15}$$

and where $\frac{\mathrm{d}\tilde{m}}{\mathrm{d}\alpha} > 0$ *for* $B''' = 0$ (*sufficient*).

Thus, the coalition countries are less [more] ambitious than the fringe countries in small [large] coalitions, in which the leakage effect outweighs [is outweighed by] the marginal emissions damage effect. The partial derivative of $\tilde{m}$ with respect to α is positive, so the respective threshold coalition $\tilde{m}$ tends to get larger as countries become more altruistic. In other words, the coalition countries tend to become less ambitious in the fight against climate change compared to the fringe countries. On the one hand, the altruism parameter ceteris paribus increases the importance of all other coalition countries' marginal emissions benefits for the optimal policy, and it increases the leakage rate to the fringe countries. On the other hand, it ceteris paribus increases the importance of all fringe countries' marginal emissions benefits for the optimal policy. Proposition 5 reveals that the former effect outweighs the latter with linear-quadratic consumption benefits.

Since $m \geq \tilde{m}$ will turn out to be the relevant coalition size and to prepare the analysis of the first stage of the Stackelberg game, we prove in Appendix A.3.4[21]

Lemma 2 (Effects of coalition size and altruism on emissions and welfare for $m \geq \tilde{m}$) *Suppose* $B''' \geq 0$ *and* $D''' \leq 0$.

- $\frac{\mathrm{d}e_f}{\mathrm{d}m} > 0, \frac{\mathrm{d}e}{\mathrm{d}m} < 0$ and $\frac{\mathrm{d}V_c}{\mathrm{d}m}, \frac{\mathrm{d}V_f}{\mathrm{d}m}, \frac{\mathrm{d}W_f}{\mathrm{d}m} > 0$ and $V(m) > V(m-1), W(m) > W(m-1)$,
- $\frac{\mathrm{d}e_f}{\mathrm{d}\alpha} < 0$ and $\frac{\mathrm{d}W}{\mathrm{d}\alpha} > 0$.

[20] For $\alpha = 0$, the existence of a threshold coalition $\tilde{m}$ has been proven by Diamantoudi and Sartzetakis (2006) for the emissions game with linear-quadratic benefits and quadratic costs, by McGinty (2020) for the abatement game with linear-quadratic benefits and quadratic costs, and by Finus et al. (2021b) for the abatement game with zero third derivatives.

[21] For $\alpha = 0$, the results of Lemma 2 have been proven by Finus et al. (2021b, Proposition 2) for the abatement game.

From the first bullet of the lemma, we get the typical results that each fringe country's emissions increase but global emissions decrease with the coalition size, so free-rider incentives tend to increase as the coalition gets larger as in the Nash game. Contrary to the Nash game, the resulting lower climate damages ensure that not only each fringe country's moral welfare but also each coalition country's moral welfare increases with the coalition size. Furthermore, each fringe country's material welfare increases with the coalition size because its consumption benefits increase and the climate damages decrease with the coalition size. Finally, Finus et al. (2021a, p. 18) prove that any Stackelberg game is superadditive, i.e. $mV_c(m) \geq (m-1)V_c(m-1) + V_f(m-1)$, because $e_c(m)$ maximizes the sum of m countries' moral welfare, taking $e_f(e_c(m))$ into account. Since each fringe country's moral welfare increases with the coalition size, superadditivity implies that global material and moral welfare also increase with the coalition size.[22]

The second bullet of the lemma reveals that each fringe country's emissions decrease with the altruism parameter and that global material welfare increases with the altruism parameter as in the Nash game. Consequently, the relative global material and moral welfare $W/W^{\mathrm{SO}} = V/V^{\mathrm{SO}}$ increase with the altruism parameter. Contrary to the Nash game, global emissions need not decrease with the altruism parameter. Finally, we prove in Appendix A.3.4 that the slope of the fringe countries' aggregate reaction function and, thus, the leakage rate to the fringe countries $|R_f'|$ increases in absolute terms with the altruism parameter for $D''' \leq 0$ and $B''' = 0$ as in the Nash game. The higher this leakage rate, the smaller ceteris paribus the coalition's ambition and the greater ceteris paribus the coalition's strategic advantage over the fringe in the Stackelberg game (Finus et al. 2021b).[23]

Now we turn to the first stage of the Stackelberg game. First note that Proposition 4 implies that all coalitions with $\tilde{\theta} \leq 1$ are externally unstable because joining this coalition then increases the respective country's moral welfare from $V_f \leq V_i^{\mathrm{BAU}}$ to $V_c > V_i^{\mathrm{BAU}}$. Consequently, global emissions are smaller and each country's moral welfare is greater at the stable Stackelberg equilibrium than at BAU, and each fringe country's welfare is greater than each coalition country's welfare. Together with Proposition 5, this gives[24,25]

Lemma 3 (Instability of small coalitions) *Suppose* $B''' \geq 0$ *and* $D''' \leq 0$. *Then, all coalitions* $m \leq \tilde{m}$ *are externally unstable, and the coalition* $m = \lfloor \tilde{m} + 1 \rfloor \geq 2$ *is internally stable.*

The coalition $m = \lfloor \tilde{m} + 1 \rfloor \geq 2$ is internally stable because leaving the coalition decreases the respective country's moral welfare from $V_c > V_i^{\mathrm{BAU}}$ to $V_f \leq V_i^{\mathrm{BAU}}$. The lemma indicates that the coalition size increases with the threshold coalition $\tilde{m}$, which in

[22] That is, $mV_c(m) + (n-m)V_f(m) = V(m) = [1 + \alpha(n-1)]W(m) \geq [1 + \alpha(n-1)]W(m-1) = V(m-1) = (m-1)V_c(m-1) + V_f(m-1) + (n-m)V_f(m-1)$ for $m \geq \tilde{m}$.

[23] However, note that the effective leakage rate $(1-\alpha)|R_f'|$ can decrease with the altruism parameter and that it is smaller with altruism than without for $\alpha \geq (n-2)/(n-1)$, $D''' = 0$ and $B''' = 0$.

[24] The function $\lfloor \cdot \rfloor$ maps its argument to the largest weakly smaller integer.

[25] For $\alpha = 0$, the results of Lemma 3 have been proven Finus et al. (2021b, Corollary 3) for the abatement game with zero third derivatives.

turn tends to increase with the altruism parameter from Proposition 5. Via this mechanism, altruism could stabilise larger coalitions.

For the detailed stability analysis, we use specification (9). We constrain the parameter space to ensure non-negative emissions for all $m \in [2, n]$, which gives an upper bound for d similar to the Nash game. In particular, we formulate the following

Assumption 2 $d \leq \bar{\bar{d}} := \dfrac{4b}{n(n-4)}$.

This assumption is necessary and sufficient for non-negative emissions for all $m \in [2, n]$ with $\alpha = 0$.[26] In Appendix A.3.5, we then prove[27]

Proposition 6 (Stability of coalitions) *Consider specification (9) with $n \geq 7$ and $d \leq \bar{\bar{d}}$.*

- Some unique coalition $m \in (\tilde{m}, \tilde{m} + 2)$ is stable.
- Some unique coalition $m \in \{2, 3\}$ is stable for $\alpha = 0$.
- The coalition size weakly increases with α.
- Some unique coalition $m \in \{2, 3, 4, 5, 6\}$ is stable for $\alpha > 0$.

Contrary to the Nash game, Proposition 6 reveals that altruism stabilises larger coalitions. However, the coalition never comprises more than six countries. More importantly, the coalition is always smaller than $m = \tilde{m} + 2$. Since the Stackelberg equilibrium and BAU coincide for $m = \tilde{m}$, the emissions-reducing and welfare-enhancing effects of the coalition size from Lemma 2 are negligible. In fact, the small coalitions stem from constraining the parameter space to ensure non-negative emissions for $m \in [2, n]$, which gives an upper bound for d/b and, thus, for $\tilde{m}$. From Proposition 5, this upper bound increases with the altruism parameter, which is the driving force for larger coalitions with than without altruism.

We use a numerical example to demonstrate there are economies in which the coalition is larger with than without altruism. Figure 3 depicts each coalition country's minimal emissions[28] (left-hand side figure) and the internal stability condition (right-hand side figure) for $m = 3$ (solid curve) and for $m = 4$ (dashed curve) dependent on α. In the numerical example, each coalition country's emissions are positive for all $m \in [2, n]$. Furthermore, $m = 3$ becomes stable for $\alpha \geq 0.223$, and $m = 4$ becomes stable for $\alpha \geq 0.839$. Thus, there are economies in which the coalition is larger with than without altruism. Finally, Fig. 4 shows that global emissions decrease and global material welfare increases with the altruism parameter in the numerical example. Furthermore, as the coalition gets larger at $\alpha = 0.223$ and at $\alpha = 0.839$, global emissions jump downwards and global material welfare jumps upwards, but these jumps are (almost) not visible.

[26] Assumption 2 is also sufficient for non-negative emissions for all $m \in [2, n]$ with $\alpha \geq 0.5$.

[27] For $\alpha = 0$, the results of Proposition 6 have been proven by Diamantoudi and Sartzetakis (2006, p. 261). For $n \in \{5, 6\}$, they show that some unique coalition $m \in (\tilde{m}, \tilde{m} + 3)$ with $m \in \{2, 3, 4\}$ $[m \in \{2, 3\}]$ is stable for $n = 5$ $[n = 6]$.

[28] Using $e_c(m(\alpha), \alpha)$ with $m(\alpha) = \arg\min e_c(m, \alpha)$.

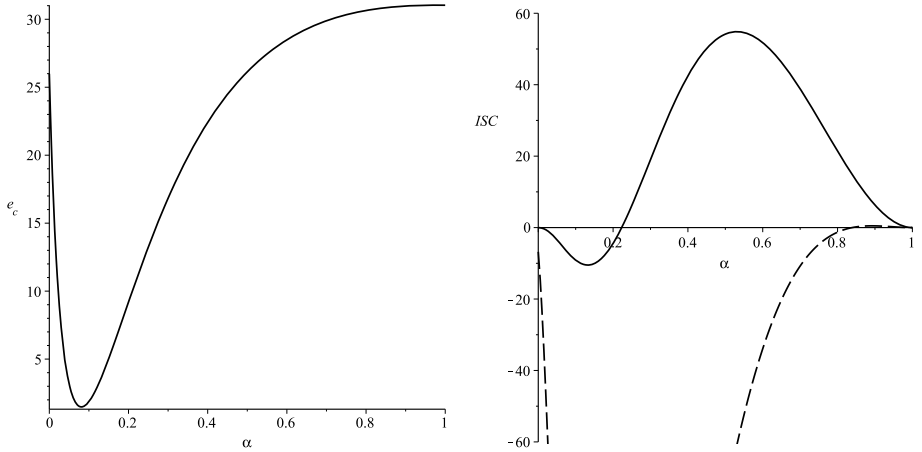

Fig. 3 Each coalition country's minimal emissions (left-hand side figure) and the internal stability condition (right-hand side figure) for $m = 3$ (solid curve) and for $m = 4$ (dashed curve) dependent on α with $n = 100, a = 100, b = 1$ and $d = 1/4500$

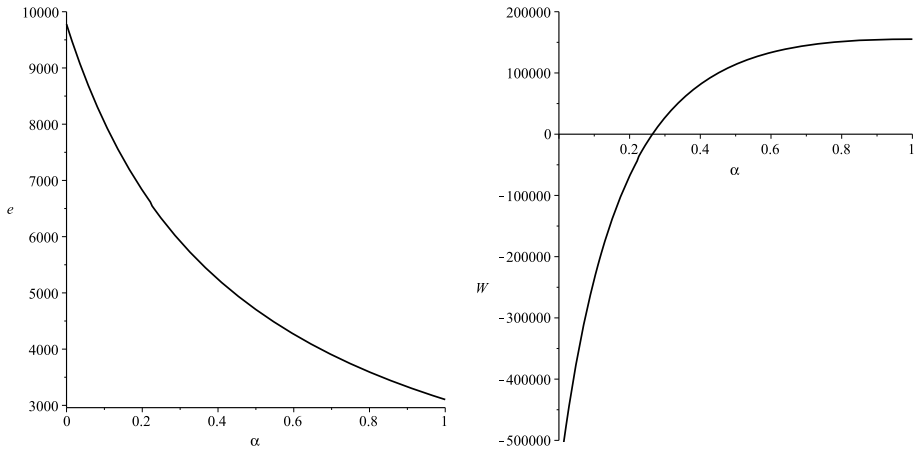

Fig. 4 Global emissions (left-hand side figure) and global material welfare (right-hand side figure) dependent on α with $n = 100, a = 100, b = 1$ and $d = 1/4500$

5 Discussion

The previous sections have shown that altruism affecting the membership decision and the policy decision leads to small (Proposition 2) or ineffective (Proposition 6) climate coalitions. By contrast, altruism affecting the membership decision only (van der Pol et al. 2012) can stabilise the grand coalition (Proposition 3).[29] Given that the effects of altruism

[29] This was shown for the Nash game, but since the stable coalitions are always weakly larger in the Stackelberg game than in the Nash game (Finus et al. 2023), this also applies to the Stackelberg game.

on the formation of climate coalitions crucially depends on its modelling, this section discusses which realms of decision making might be influenced by social preferences in general.[30]

We start by looking at what assumptions the literature on self-enforcing international environmental agreements with social preferences makes and how it justifies them. First, the literature incorporating inequality aversion (Lange and Vogt 2003; Lange 2006; Vogt 2016; Rogna and Vogt 2022) considers social preferences at both stages of the game and argues that governments interested in re-election must take (median) voters' fairness preferences into account in national policy and international negotiations. Second, the literature incorporating reciprocal fairness (Grüning and Peters 2010; Nyborg 2018) also considers social preferences at both stages of the game.[31] Nyborg (2018, p. 707) argues that although groups may act differently than individuals, policy makers and treaty negotiators tend to be reciprocal when the general population is reciprocal, and tend to act reciprocally when the median voter is reciprocal. Third, the literature incorporating Kantian ethics (Eichner and Pethig 2022; Ulph and Ulph 2023) allows for different moral behaviour at the two stages of the game. Nevertheless, Eichner and Pethig (2022, pp. 18–19) argue that moral behaviour at just one stage of the game or different moral behaviour at the two stages of the game appears to be implausible, and Ulph and Ulph (2023, p. 12) "recognise that there is a strong argument that an agent should take the same moral stance to all decisions." However, the latter argue that governments decide on coalition membership, while both governments and individuals decide on domestic emissions through public policy and private behaviour, respectively, such that the "decisions involve somewhat different agents", which might explain different moral behaviour at the two stages of the game. Finally, van der Pol et al. (2012, p. 114) argue that "agents may hold different preferences when acting in different social situations, for example as consumers or as citizens." They then distinguish between the decision about the technology employed and the domestic regulations adopted of a homo economicus with "personal well-being functions" (Nyborg 2000, p. 305), and the decision about the membership in the coalition of a homo politicus with "subjective social welfare functions" (Nyborg 2000, p. 305).

To sum up, most of the literature assumes that individual social preferences affect both stages of the game via the median voter (or because policy makers and treaty negotiators tend to have the same social preferences as the general population). On the other hand, taken together, van der Pol et al. (2012) and Ulph and Ulph (2023) provide good arguments that social preferences could be different at the two stages of the game: If there is a distinction between homo economicus and homo politicus and if homo economicus can influence domestic emissions, then social preferences might be more pronounced at the membership stage than at the policy stage. However, note that the cited literature (and the present paper) abstracts from individual decisions and assumes that countries or governments decide on both coalition membership and domestic emissions. In this case, all decisions are made by citizens rather than by consumers, and there should be no qualitative difference between social preferences at the two stages of the game.

[30] We owe this section to an anonymous reviewer who suggested to discuss which realms of decision making might be influenced by altruistic preferences.

[31] In Buchholz et al. (2018) countries are reciprocal when they decide on their membership in the coalition, but the emissions policy inside and outside the coalition is exogenously given.

However, strategic delegation could be another argument for different social preferences at different game stages. The literature on strategic delegation shows that strategic voters elect policy makers who care less about global public goods when policy makers bargain over their provision. Thereby, Buchholz et al. (2005) and Loeper (2017) assume that policy makers decide on the cooperative bargaining outcome and on the non-cooperative bargaining default (strong delegation), while Segendorff (1998) and Graziosi (2009) also consider policy makers deciding on the cooperative bargaining outcome but median voters deciding on the non-cooperative bargaining default (weak delegation), which leads to different preferences at different game stages.[32] Spycher and Winkler (2022) introduce strategic delegation into the standard two-stage game of self-enforcing international environmental agreements. They distinguish between weak delegation, i.e. elected policy makers decide on coalition membership but median voters decide on emissions policy, and strong delegation, i.e. elected policy makers decide at both stages of the game. While weak delegation does not increase coalition size but does increase global emissions, strong delegation can stabilise the grand coalition and thereby bring about the social optimum. These results suggest that strategic delegation can pay off. In this context, Lange and Schwirplies (2017) argue that there is indeed strategic delegation in climate policy because the social preferences of climate negotiators and the general population differ in that the former are more likely to support burden-sharing rules with low economic costs for their regions than the latter.

6 Extensions

This section performs two model extensions. The first subsection shows that a small degree of community altruism can stabilise the grand coalition. The second subsection reveals that altruism can stabilise coalitions of two countries but tends to destabilise coalitions of three or more countries with linear climate damages and asymmetric countries.[33]

6.1 Community Altruism

In this subsection, we consider community altruism. In particular, we distinguish between out-group altruism α and in-group altruism $\beta > \alpha$. The psychological literature has developed and tested two theories in particular for the preference of in-group members over out-group members (Balliet et al. 2014): The social identity theory assumes that individuals identify themselves with their memberships in social groups (Tajfel et al. 1979), while the theory of bounded generalized reciprocity assumes that groups contain individuals with cooperative reputations, which induces indirect reciprocity (Yamagishi and Kiyonari 1999). Cheikbossian (2021a) shows that in-group altruism is evolutionary stable, Cheikbossian (2021b) finds that a combination of in-group altruism and out-group altruism can be evolutionary stable, and Balliet et al. (2014) summarize the significant evidence for in-group altruism (and against out-group spite) in psychological experiments.

[32] While Buchholz et al. (2005) and Graziosi (2009) find that strategic voting cancels the gains of international cooperation, Segendorff (1998) finds that weak [strong] delegation increases [decreases] the gains of international cooperation, and Loeper (2017) finds that the results depend on the type of global public goods.

[33] We owe this section to an anonymous reviewer who suggested to extend the analysis to community altruism and asymmetric countries.

In the following, we first analyse community altruism in the Nash game then report the effects of community altruism in the Stackelberg game. In the Nash game of coalition formation, each fringe country does not belong to a group, such that its moral welfare (1) and its first-order condition (4) do not change. By contrast, each coalition country belongs to the climate coalition, such that its moral welfare becomes

$$\begin{aligned}\sum_{i\in M} V_i &= \sum_{i\in M}\left[W_i + \beta\sum_{j\in M\setminus i} W_j + \alpha\sum_{j\in N\setminus M} W_j\right]\\ &= [1-\alpha+(\beta-\alpha)(m-1)]\sum_{i\in M} W_i + \alpha m W.\end{aligned} \tag{16}$$

Thus, community altruism ceteris paribus increases the weight of the coalition's material welfare. Rearranging the corresponding first-order condition yields[34]

$$B'(e_c) = \frac{1+\alpha(n-m)+\beta(m-1)}{1+\beta(m-1)} m D'(e) \leq n D'(e), \tag{17}$$

where the numerator reflects the altruistic preferences of the respective coalition country for all other countries' climate damages, and the denominator reflects the altruistic preferences of all other coalition countries for the respective coalition country's consumption benefits. Community altruism reduces the relative importance of the fringe countries' climate damages, which ceteris paribus increases the coalition's emissions. However, from (4) and (17), we infer

$$\frac{B'(e_c)}{B'(e_f)} = \frac{m}{1+\alpha(n-1)} \frac{1+\alpha(n-m)+\beta(m-1)}{1+\beta(m-1)} \in (1, m]. \tag{18}$$

Consequently, each fringe country's emissions are still greater than each coalition country's emissions. In Online Appendix B.1.1, we prove that the first bullet and the third bullet of Proposition 1 also hold with community altruism: Each coalition country consumes less than at BAU, which reduces global emissions and, thus, marginal climate damages, such that each fringe country consumes more than at BAU. Consequently, the material welfare of each fringe country is greater than that of each coalition country and greater than that at BAU. However, and in contrast to the case without community altruism, the moral welfare of each coalition country can be greater than that of each fringe country if the material welfare of each coalition country is positive, because then community altruism ceteris paribus increases the former but does not affect the latter. In particular, we have $V_f - V_c = (1-\alpha)(W_f - W_c) - (\beta-\alpha)(m-1)W_c$, which ceteris paribus decreases with the degree of community altruism and with the coalition size if $W_c > 0$. This welfare effect of community altruism can reduce the free-rider incentives.

To analyse the policy effect of community altruism, we prove in Online Appendix B.1.1[35]

Lemma 4 (Effects of community altruism on emissions and welfare)

[34] The second-order conditions are fulfilled.

[35] Furthermore, we there prove that Lemma 1 also holds with community altruism.

- $\frac{\mathrm{d}e_f}{\mathrm{d}\beta} < 0, \frac{\mathrm{d}e_c}{\mathrm{d}\beta} > 0$ and $\frac{\mathrm{d}e}{\mathrm{d}\beta} > 0$,
- $\frac{\mathrm{d}W_f}{\mathrm{d}\beta} < 0$ and $\frac{\mathrm{d}W_c}{\mathrm{d}\beta} > 0$,
- $\frac{\mathrm{d}V_c}{\mathrm{d}\beta} > 0$ and $\frac{\mathrm{d}(V_f - V_c)}{\mathrm{d}\beta} < 0$ for $W_c \geq 0$ (sufficient).

Ceteris paribus, community altruism does not affect each fringe country's consumption, see (4), but increases each coalition country's consumption, see (17). Thus, e_c increases, which increases global emissions and, thus, marginal climate damages, such that e_f decreases. Consequently, each fringe country's material welfare decreases, while the increase in each coalition country's consumption benefits outweighs the increase in the climate damages, such that its material welfare increases. This policy effect of community altruism raises each coalition country's moral welfare and reduces the difference between the moral welfare of each fringe country and each coalition country if $W_c \geq 0$. Furthermore, it reduces each fringe country's moral welfare $V_f = (1-\alpha)W_f + \alpha W$ if α is sufficiently small, such that the decrease in W_f outweighs the potential increase in W.

Taken together, the sign of each coalition country's material welfare plays an important role for the effects of community altruism on the free-rider incentives. This also becomes clear when we look at the (internal) stability condition of the grand coalition:

$$\begin{aligned} &V_c(n) - V_f(n-1) \\ &= (1-\beta)W_c(n) + \beta W(n) - \big[(1-\alpha)W_f(n-1) + \alpha W(n-1)\big] \\ &= W(n) - (1-\beta)(n-1)W_c(n) - \big[W(n-1) - (1-\alpha)(n-1)W_c(n-1)\big]. \end{aligned} \tag{19}$$

First note that the grand coalition maximizes global material welfare from $V_c(n) = [1+\beta(n-1)]W_c(n)$ as in the case without community altruism. Now consider $\beta = 1$. Then, $V_c(n)$ is maximized global material welfare $W(n)$, and $V_f(n-1)$ is non-maximized global material welfare $W(n-1)$ minus the coalition's material welfare weighted by $1-\alpha$. Thus, the grand coalition is stable if $W_c(n-1) \geq 0$. By contrast, if $W_c(n-1) < 0$, the grand coalition need not be stable. For example, consider $\alpha = 0$. Then, $V_f(n-1)$ is equal to the fringe country's material welfare $W_f(n-1)$, and this material welfare can exceed maximized global welfare $W(n)$ if $W_c(n-1) < 0$. However, it can be shown that $W_c(n-1) > 0$ holds with specification (9) when there is a sufficiently large number of countries ($n \geq 9$). Then, the emissions policy of one fringe country is relatively unimportant for the material welfare of many coalition countries. In this case, the grand coalition is stable for $\beta = 1$.

Note that $W_c(n)$ is always non-negative because the grand coalition could choose $e_c = 0$ and, thus, $B(0) \geq 0$ and $D(0) = 0$. Consequently, in-group altruism increases each country's moral welfare in the grand coalition. Furthermore, it reduces each fringe country's material welfare, but raises each coalition country's material welfare with any other coalition, such that the effect on each fringe country's moral welfare is in general ambiguous. However, in Online Appendix B.1.1 we prove that this effect is definitely negative if there is only one fringe country. Then, in-group altruism increases the emissions of so many coalition countries that the corresponding increase in climate damages and decrease in the fringe country's consumption benefits outweigh the increase in the coalition country's consumption benefits. To sum up, in-group altruism increases the stability of the grand coalition.

Finally, out-group altruism does not affect each country's moral welfare in the grand coalition, but it reduces global emissions and increases global

material welfare with any other coalition, see Lemma 1. Thus, it also tends to increase $V_f(n-1) = W_f(n-1) + \alpha(n-1)W_c(n-1)$ if $W_c(n-1) > 0$. Consequently, it can be shown that out-group altruism increases the fringe country's moral welfare with specification (9) when there is a sufficiently large number of countries ($n \geq 10$). In this case, out-group altruism decreases the stability of the grand coalition. We prove our results in Online Appendix B.1.1 and summarize them in

Proposition 7 (Stability of grand coalition with community altruism) *Consider specification* (9).

- The internal stability condition of the grand coalition increases with β, and it decreases with α for $n \geq 10$ and $d \leq \bar{d}$ (sufficient).
- The grand coalition is stable for $\beta = 1$, $n \geq 6$ and $d \leq \bar{d}$ (sufficient).

Thus, community altruism can stabilise the grand coalition and thereby bring about the social optimum. Figure 5 depicts thresholds of $\beta - \alpha$ for the grand coalition to be stable dependent on α in the Nash game (left-hand side figure) for different values of the climate damage parameter d.[36] When this parameter is small, emissions policy is determined mainly by marginal emissions benefits rather than marginal climate damages, so that the difference between the emissions and therefore the material welfare of each fringe country and each coalition country is small, which reduces free-rider incentives and stabilises the grand coalition. Figure 5 shows that the grand coalition can be stable even with a small difference between in-group altruism and out-group altruism of less than 0.03.[37] Note that community altruism does not change the grand coalition's emissions policy, so its stabilizing effect relies on the worse emissions policy for the final fringe country (policy effect) and on the increased moral welfare of each coalition country for a given emissions policy (welfare effect).[38]

Finally, we report the effects of community altruism in the Stackelberg game. Since the results are simply a combination of those in the Stackelberg game without community altruism and those in the Nash game with community altruism, we delegate the full analysis to Online Appendix B.1.2. First, there is a unique threshold coalition $\hat{m} \in (1, n)$ where the Stackelberg equilibrium and BAU coincide in terms of emissions and material welfare. If $W_i^{\text{BAU}} \geq 0$, then $V_c(m) \geq V_i^{\text{BAU}}$ and $V_f(m) \leq V_i^{\text{BAU}}$ for $m \leq \hat{m}$, such that all coalition $m \leq \hat{m}$ are externally unstable, and the coalition $m = \lfloor \hat{m} + 1 \rfloor \geq 2$ is internally stable as in the case without community altruism.[39] Second, community altruism increases each coalition country's emissions and global emissions, and it decreases each fringe country's

[36] For $n = 100$ and $d/b > 3.75/10{,}000$, each coalition country's minimal emissions become negative for certain combinations of α and β.

[37] Coincidentally, this is consistent with the numerical analysis of van der Pol et al. (2012), who find that the grand coalition is stable for $\alpha = 0$ and $\beta \geq 0.03$ or $\alpha \leq 0.024$ and $\beta = 0.06$ and unstable for $\alpha \geq 0.036$ and $\beta = 0.06$.

[38] Finally, we ran several examples to find other possible stable coalition sizes. There are economies in which no coalition is stable, only the coalition $m = 2$ is stable, only the grand coalition is stable, or some coalition $m \geq 2$ and the grand coalition are stable. We did not find an example in which the coalition $m = 2$ is stable in the case without community altruism and no coalition is stable in the case with community altruism. However, there are economies in which only the coalition $m = 2$ is stable in both cases, and in which global moral welfare is smaller with than without community altruism due to the higher emissions damages.

[39] By contrast, if $W_i^{\text{BAU}} < 0$, then $V_c(\hat{m}) < V_i^{\text{BAU}}$ and $V_f(\hat{m}) = V_i^{\text{BAU}}$, such that the coalition $m = \hat{m} - 1$ is externally stable.

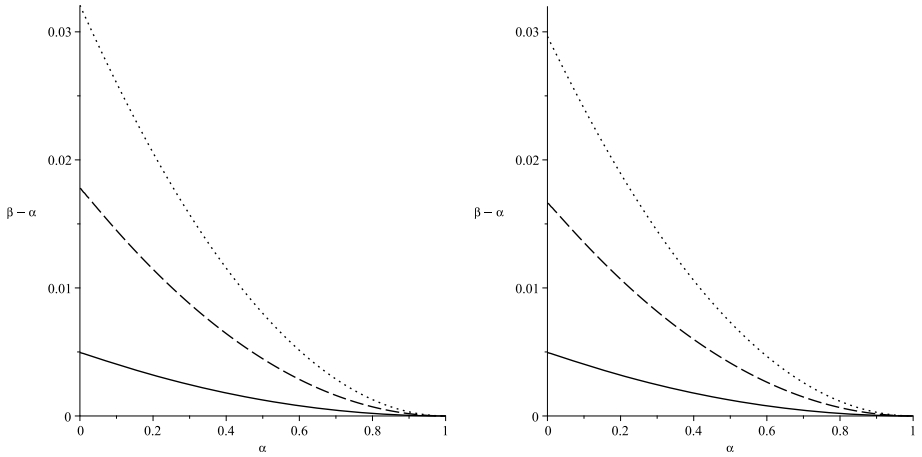

Fig. 5 Thresholds of $\beta - \alpha$ for the grand coalition to be stable dependent on α in the Nash game (left-hand side figure) and in the Stackelberg game (right-hand side figure) with $n = 100, a = 100, b = 1$ and $d = 1/10{,}000$ (solid curve), $d = 2.375/10{,}000$ (dashed curve) and $d = 3.75/10{,}000$ (dotted curve)

emissions as in the Nash game. Consequently, $\hat{m}$ increases with β, which can stabilise larger coalitions. Furthermore, community altruism decreases each fringe country's moral welfare, and it increases each coalition country's moral welfare if $W_c \geq 0$ as in the Nash game, which then stabilises larger coalitions. In particular, in-group [out-group] altruism increases [decreases] the internal stability condition of the grand coalition [when there is a sufficiently large number of countries] as in the Nash game. Furthermore, it can be shown that the condition for the grand coalition to be stable is laxer in the Stackelberg game than in the Nash game. However, Fig. 5 shows that the thresholds of $\beta - \alpha$ for the grand coalition to be stable dependent on α in the Stackelberg game (right-hand side figure) are close to those in the Nash game (left-hand side figure).

6.2 Asymmetric Countries

In this subsection, we consider asymmetric countries in terms of consumption benefits $B_i(e_i)$ and climate damages $D_i(e)$. With convex climate damages $D_i'' > 0$ (or abatement benefits), the results of previous literature on coalition stability with asymmetric countries are based on numerical analyses (Barrett 1997; Botteon and Carraro 2001; McGinty 2007; Bakalova and Eyckmans 2019; McGinty 2020). In order to obtain analytical results, we thus rely on linear climate damages $D_i'' = 0$ in this subsection. This implies that each fringe country has a dominant strategy, such that the Nash game and the Stackelberg game coincide. Furthermore, we focus on the case without transfers.

In such a framework with two types of countries, Fuentes-Albero and Rubio (2010) show that a coalition of at most three countries is stable if either climate damages or emissions benefits are symmetric. Pavlova and De Zeeuw (2013) confirm this result and show that the same holds if climate damages and emissions benefits are positively correlated. By contrast, if climate damages and abatement costs are negatively correlated, a

coalition of all countries with low climate damages and two countries with high climate damages can be stable. Finus and McGinty (2019) extend this analysis by allowing for any type of countries. They show that the grand coalition can be stable if climate damages and abatement costs are negatively correlated.

Following van der Pol et al. (2012), we consider a unique altruism parameter to derive the general effects of altruism on the formation of climate coalitions. Furthermore, we abstract from community altruism to check the robustness of our main model with linear climate damages and asymmetric countries. With a unique altruism parameter and without community altruism, each country's moral welfare function is still given by (1), and the global moral welfare function is still given by (2). Consequently, the socially optimal emissions always maximize global material welfare, and the individually optimal emissions maximize global material welfare if and only if $\alpha = 1$. In Appendix A.1, we prove that global emissions decrease and global material welfare increases with the altruism parameter in the individually optimal solution as in the main model. Consequently, the relative global emissions $e^{\mathrm{BAU}}/e^{\mathrm{SO}}$ decrease and the relative global material and moral welfare $W^{\mathrm{BAU}}/W^{\mathrm{SO}} = V^{\mathrm{BAU}}/V^{\mathrm{SO}}$ increase with the altruism parameter.

In the coalition formation game, each fringe country's moral welfare is still given by (1), and the sum of the coalition countries' moral welfare is still given by (3). Rearranging the corresponding first-order conditions yields

$$B_i'(e_i) = (1-\alpha)D_i'(e) + \alpha \sum_{j\in N} D_j'(e) \leq \sum_{j\in N} D_j'(e), \quad \forall i \notin M, \tag{20}$$

$$\begin{aligned} B_i'(e_i) &= \frac{(1-\alpha)\sum_{j\in M} D_j'(e) + \alpha m \sum_{j\in N} D_j'(e)}{1+\alpha(m-1)} \\ &= (1-\alpha)D_i'(e) + \alpha \sum_{j\in N} D_j'(e) \\ &\quad + (1-\alpha)\frac{\sum_{j\in M\setminus i} D_j'(e) + \alpha(m-1)\sum_{j\in N\setminus i} D_j'(e)}{1+\alpha(m-1)} \leq \sum_{j\in N} D_j'(e), \quad \forall i \in M. \end{aligned} \tag{21}$$

Consequently, considering two ex ante identical countries, the emissions of the country outside the coalition are greater than those of the country inside the coalition, which means that also the material and moral welfare of the country outside the coalition are greater than those of the country inside the coalition. In Online AppendixB.2.1, we prove[40]

Proposition 8 (Comparison of Nash equilibrium and BAU with asymmetric countries)

- $e_i < e_i^{\mathrm{BAU}}$ for all $i \in M$, $e_i = e_i^{\mathrm{BAU}}$ for all $i \notin M$ and $e < e^{\mathrm{BAU}}$,
- $V_i > V_i^{\mathrm{BAU}}$ for all $i \notin M$ and $V > V^{\mathrm{BAU}}$,
- $W_i > W_i^{\mathrm{BAU}}$ for all $i \notin M$ and $W > W^{\mathrm{BAU}}$.

(21) implies that the coalition countries are ceteris paribus more ambitious in the fight against climate change than at BAU. This results in smaller emissions inside the coalition,

[40] For $\alpha = 0$, the results of Proposition 8 have been proven by Finus and McGinty (2019, p. 544) for the abatement game with linear benefits and quadratic costs.

while the fringe countries' dominant strategies imply that the emissions outside the coalition do not change, such that global emissions are smaller than at BAU. Smaller global emissions and constant fringe country's emissions imply $W_i > W_i^{\text{BAU}}$ for all $i \notin M$. The remaining results arise from the superadditivity of the game, i.e. $\sum_{j\in M} V_j \geq \sum_{j\in M} V_j^{\text{BAU}}$: Superadditivity implies $W > W^{\text{BAU}}$, which in turn implies $V_i > V_i^{\text{BAU}}$ for all $i \notin M$ and $V > V^{\text{BAU}}$.

To prepare the analysis of the first stage of the Nash game, we prove in Online Appendix B.2.2[41]

Lemma 5 (Effects of coalition size and altruism on emissions and welfare with asymmetric countries)

- If another country joins the coalition, then each coalition country's emissions decrease, each fringe country's emissions do not change and global emissions decrease. Furthermore, each fringe country's material and moral welfare increase, and global material and moral welfare increase.
- $\frac{\mathrm{d}e_i}{\mathrm{d}\alpha} < 0$ for all $i \in N$ and $\frac{\mathrm{d}W}{\mathrm{d}\alpha} > 0$.

The first bullet of the lemma reveals that the comparison between Nash equilibrium and BAU from Proposition 8 can be transferred to the comparison between the equilibrium with some coalition M and the equilibrium with some smaller coalition $M\backslash i$: The larger the coalition, the more it fights against climate change, which leads to smaller emissions of the original coalition members, to smaller emissions of the new coalition member and, thus, to smaller global emissions. The material welfare of each fringe country increases, because its consumption benefits do not change, but its climate damages decrease. Superadditivity implies that global welfare increases, which in turn implies that the moral welfare of each fringe country and global moral welfare increase.

The second bullet of the lemma reveals that each country's emissions decrease and global material welfare increases with the altruism parameter. Consequently, the relative global emissions e/e^{SO} decrease and the relative global material and moral welfare $W/W^{\text{SO}} = V/V^{\text{SO}}$ increase with the altruism parameter as in the main model with symmetric countries and convex climate damages.

For the first stage of the game, we use the following specification

$$B_i(e_i) = a_i e_i - \frac{b_i}{2} e_i^2, \quad D_i(e) = d_i e, \quad \text{with } a_i, b_i, d_i > 0. \tag{22}$$

In Online Appendix B.2.3, we then prove

Proposition 9 (Stability of coalitions with asymmetric countries) *Consider the specification* (22).

- Suppose $b_i = b$ and $d_i = d$ for all $i \in N$. Then, any coalition with three members is stable for $\alpha = 0$, and any coalition with two members is stable for $\alpha > 0$.

[41] For $\alpha = 0$, the results of Lemma 5 have been proven by Finus and McGinty (2019, p. 544) for the abatement game with linear benefits and quadratic costs.

- Suppose $\{b_1 < b_2 < \cdots < b_n$ and $d_1 \leq d_2 \leq \cdots \leq d_n\}$ or $\{b_1 \leq b_2 \leq \cdots \leq b_n$ and $d_1 < d_2 < \cdots < d_n\}$. Then, only coalitions with two members can be stable, and the condition for at least one coalition to be stable is tighter for $\alpha = 0$ than for $\alpha > 0$.
- Suppose $b_1 \leq b_2 \leq \cdots \leq b_n$ and $d_1 \geq d_2 \geq \cdots \geq d_n$. Then, any coalition can be stable, and the condition for the grand coalition to be stable is laxer for $\alpha = 0$ than for $\alpha > 0$.

For $\alpha = 0$, Proposition 9 replicates the results of Finus and McGinty (2019): With identical b_i and d_i parameters, a coalition of three countries is stable. With a positive covariance between these parameters, either a coalition of two countries or no coalition is stable. The coalition becomes smaller because countries with small b_i and d_i parameters profit greatly from additional emissions benefits and suffer little from additional climate damages when they leave a coalition.[42] With a negative covariance, countries with small b_i and large d_i parameters and countries with large b_i and small d_i parameters can mutually benefit from a large coalition, which can even stabilise the grand coalition.

For $\alpha > 0$, we get mixed results. With identical b_i and d_i parameters, the negative effect of altruism is even stronger than with quadratic climate damages, since the coalition is always smaller for $\alpha > 0$ than for $\alpha = 0$ with linear climate damages. In Online Appendix B.2.4, we prove that the effects of altruism on the internal stability condition are comparable to those of the main model: The direct effect is positive, the policy effect inside the coalition is positive, but the policy effect outside the coalition is also positive and predominates, such that staying in the coalition becomes less important as countries become more altruistic. With individual b_i and d_i parameters, altruism has two effects on the internal stability condition: First, it reduces the effective variance of the d_i parameters, because each coalition country accounts for the climate damages of all fringe countries. Second, it increases the incentive to stay in small coalitions $|M| = 2$, but reduces the incentive to stay in large coalitions $|M| \geq 3$. Both effects stabilise $|M| = 2$ with a positive covariance between b_i and d_i, but tend to destabilise $|M| \geq 3$ with a negative covariance between these parameters. In particular, the condition for the grand coalition to be stable is laxer for $\alpha = 0$ than for $\alpha > 0$.

Finally, to further check the robustness of our main model, we also analyse the second stage of the Nash game with convex climate damages and asymmetric countries in Online Appendix B.2. We find that global emissions decrease and each fringe country's emissions and material welfare increase if another country joins the coalition as in the main model. Consequently, global emissions are smaller and each fringe country's emissions and material welfare are greater than at BAU. Furthermore, global emissions decrease with the altruism parameter as in the main model. However, in contrast to the main model, some fringe country's emissions could increase with the altruism parameter. For example, suppose that only one fringe country faces climate damages and countries become more altruistic. Then, all other countries consume less (because this benefits the respective fringe country), which reduces global emissions and, thus, marginal climate damages, such that the respective fringe country consumes more (because this does not harm the other countries). The effects of altruism on global material welfare also become ambiguous. However, if countries only differ in their emissions benefits and not in their climate damages, then each fringe country's emissions decrease and global material welfare increases with the altruism parameter as in the main model.

[42] In Online Appendix B.2.3, we prove that $n \in M$ and $2d_{n-1}^2 b_{n-1} \geq d_n^2 b_n$ is necessary and sufficient for at least one coalition of two countries to be stable.

7 Conclusion

This paper analyses the effects of altruism on the formation of climate coalitions in the canonical two-stage game of self-enforcing international environmental agreements. Thereby, altruism implies that each country values, to some extent, every other country's welfare when deciding on its coalition membership and emissions policy. In the Nash [Stackelberg] game with linear-quadratic emissions benefits and quadratic emissions damages, altruism weakly decreases [increases] the coalition size. However, the coalition never comprises more than six countries, and the corresponding global emissions and global welfare are close to the non-cooperative values. Nevertheless, altruism reduces global emissions and raises global welfare by narrowing the gap between the individually optimal values and the socially optimal values. Thus, altruism affecting the membership decision and the policy decision appears to be more of a substitute than a complement for large climate coalitions. Consequently, altruism may help explain why countries are willing to internalize their climate externalities onto other countries, but are unwilling to conclude a large and effective climate agreement.

We find that these results crucially depend on the modelling of altruism: If altruism affects the membership decision but not the policy decision (van der Pol et al. 2012), countries stay in the coalition to avoid worse policy outcomes from smaller coalitions, which can stabilise large coalitions. If each coalition country is more altruistic toward other coalition countries than toward fringe countries (community altruism), coalition countries achieve greater moral welfare and become less ambitious in the fight against climate change, which reduces the free-rider incentives and can stabilise the grand coalition. Finally, altruism can stabilise small coalitions but destabilises large coalitions with linear climate damages and asymmetric countries.

Our analysis can be extended in several directions. For example, one could analyse the optimal strategic delegation of each country's principal to a country's agent with different altruistic preferences between the first and the second stage of the game (Spycher and Winkler 2022). Furthermore, it may be interesting to replace the assumption of pure altruism with the assumption of paternalistic or impure altruism. In the first case, one could consider different altruistic parameters for other countries' consumption benefits and climate damages. In the second case, one could add warm-glow transfers between countries at a third stage of the game. Finally, the results at the first stage of the Nash game and the Stackelberg game depend on the functional forms of the benefit function and the damage function, such that it may be interesting to replace our linear-quadratic specification with, e.g., an isoelastic specification (Nkuiya 2020). These issues are beyond the scope of the present paper but may represent interesting and important tasks for future research.

Appendix

Business-as-Usual Scenario

The first-order condition of (1) reads

$$B_i' - D_i' - \alpha \sum_{j \in N \setminus i} D_j' = 0, \tag{A.1}$$

and the second-order condition reads

$$B_i'' - D_i'' - \alpha \sum_{j \in N \setminus i} D_j'' < 0, \tag{A.2}$$

which is fulfilled. Differentiating (A.1) with respect to α yields

$$\begin{aligned} & B_i'' \frac{\mathrm{d}e_i}{\mathrm{d}\alpha} - [D_i'' + \alpha \sum_{j \in N \setminus i} D_j''] \frac{\mathrm{d}e}{\mathrm{d}\alpha} - \sum_{j \in N \setminus i} D_j' = 0 \\ \Leftrightarrow \quad & \frac{\mathrm{d}e_i}{\mathrm{d}\alpha} = [D_i'' + \alpha \sum_{j \in N \setminus i} D_j'']/B_i'' \frac{\mathrm{d}e}{\mathrm{d}\alpha} + \sum_{j \in N \setminus i} D_j'/B_i'', \end{aligned} \tag{A.3}$$

which is negative if $D_i'' = 0$ for all $i \in N$. Taking the sum over all $i \in N$ and rearranging yields

$$\begin{aligned} & \sum_{i \in N} \frac{\mathrm{d}e_i}{\mathrm{d}\alpha} = \sum_{i \in N} [D_i'' + \alpha \sum_{j \in N \setminus i} D_j'']/B_i'' \frac{\mathrm{d}e}{\mathrm{d}\alpha} + \sum_{i \in N} \sum_{j \in N \setminus i} D_j'/B_i'' \\ \Leftrightarrow \quad & \frac{\mathrm{d}e}{\mathrm{d}\alpha} = \frac{\sum_{i \in N} \sum_{j \in N \setminus i} D_j'/B_i''}{1 - \sum_{i \in N} [D_i'' + \alpha \sum_{j \in N \setminus i} D_j'']/B_i''} < 0. \end{aligned} \tag{A.4}$$

Substituting into (A.3) yields

$$\begin{aligned} \frac{\mathrm{d}e_i}{\mathrm{d}\alpha} &= [D_i'' + \alpha \sum_{j \in N \setminus i} D_j'']/B_i'' \frac{\sum_{i \in N} \sum_{j \in N \setminus i} D_j'/B_i''}{1 - \sum_{i \in N} [D_i'' + \alpha \sum_{j \in N \setminus i} D_j'']/B_i''} + \sum_{j \in N \setminus i} D_j'/B_i'' \\ &= \frac{[D_i'' + \alpha \sum_{j \in N \setminus i} D_j''] \sum_{i \in N} \sum_{j \in N \setminus i} D_j'/B_i'' - \sum_{i \in N} [D_i'' + \alpha \sum_{j \in N \setminus i} D_j'']/B_i'' \sum_{j \in N \setminus i} D_j'}{B_i'' [1 - \sum_{i \in N} [D_i'' + \alpha \sum_{j \in N \setminus i} D_j'']/B_i'']} \\ &\quad + \frac{\sum_{j \in N \setminus i} D_j'/B_i''}{1 - \sum_{i \in N} [D_i'' + \alpha \sum_{j \in N \setminus i} D_j'']/B_i''}, \end{aligned} \tag{A.5}$$

which is negative if $D_i'' = 0$ or $D_i = D$ for all $i \in N$. Finally, differentiating W with respect to α and using (A.1) yields

$$\begin{aligned} \frac{\partial W}{\partial \alpha} &= \sum_{i \in N} B_i' \frac{\mathrm{d}e_i}{\mathrm{d}\alpha} - \sum_{i \in N} D_i' \frac{\mathrm{d}e}{\mathrm{d}\alpha} = \sum_{i \in N} [D_i' + \alpha \sum_{j \in N \setminus i} D_j'] \frac{\mathrm{d}e_i}{\mathrm{d}\alpha} - \sum_{i \in N} D_i' \frac{\mathrm{d}e}{\mathrm{d}\alpha} \\ &= -(1 - \alpha) \sum_{i \in N} D_i' [\frac{\mathrm{d}e}{\mathrm{d}\alpha} - \frac{\mathrm{d}e_i}{\mathrm{d}\alpha}], \end{aligned} \tag{A.6}$$

which is positive if $D_i'' = 0$ or $D_i = D$ for all $i \in N$.

Nash Game

Proof of Proposition 1

From (4), (5) and $\Theta := \frac{m}{1+\alpha(m-1)}$, the equilibrium is characterised by

$$B'(e_f) = [1 + \alpha(n-1)]D', \tag{A.7}$$

$$B'(e_c) = [1 + \alpha(n-1)]\Theta D', \tag{A.8}$$

$$e = me_c + (n-m)e_f. \tag{A.9}$$

First differentiating (A.7), (A.8) and (A.9) with respect to Θ yields

$$B''(e_f)\frac{\mathrm{d}e_f}{\mathrm{d}\Theta} = [1 + \alpha(n-1)]D''\frac{\mathrm{d}e}{\mathrm{d}\Theta}, \tag{A.10}$$

$$B''(e_c)\frac{\mathrm{d}e_c}{\mathrm{d}\Theta} = [1 + \alpha(n-1)]\left[\Theta D''\frac{\mathrm{d}e}{\mathrm{d}\Theta} + D'\right], \tag{A.11}$$

$$\frac{\mathrm{d}e}{\mathrm{d}\Theta} = m\frac{\mathrm{d}e_c}{\mathrm{d}\Theta} + (n-m)\frac{\mathrm{d}e_f}{\mathrm{d}\Theta}. \tag{A.12}$$

Solving for $\frac{\mathrm{d}e}{\mathrm{d}\Theta}$, $\frac{\mathrm{d}e_f}{\mathrm{d}\Theta}$ and $\frac{\mathrm{d}e_c}{\mathrm{d}\Theta}$ yields

$$\frac{\mathrm{d}e_f}{\mathrm{d}\Theta} = \frac{m[1 + \alpha(n-1)]^2 D''D'}{B''(e_c)B''(e_f) - [1 + \alpha(n-1)][(n-m)B''(e_c) + m\Theta B''(e_f)]D''} > 0, \tag{A.13}$$

$$\frac{\mathrm{d}e_c}{\mathrm{d}\Theta} = -\frac{[1 + \alpha(n-1)]\{(n-m)[1 + \alpha(n-1)]D'' - B''(e_f)\}D'}{B''(e_c)B''(e_f) - [1 + \alpha(n-1)][(n-m)B''(e_c) + m\Theta B''(e_f)]D''} < 0, \tag{A.14}$$

$$\frac{\mathrm{d}e}{\mathrm{d}\Theta} = \frac{m[1 + \alpha(n-1)]B''(e_f)D'}{B''(e_c)B''(e_f) - [1 + \alpha(n-1)][(n-m)B''(e_c) + m\Theta B''(e_f)]D''} < 0. \tag{A.15}$$

Note that $\Theta = 1 \iff e_f = e_c = e_i^{\mathrm{BAU}}$. Thus, $\Theta > 1 \implies e_f > e_i^{\mathrm{BAU}} > e_c \wedge e^{\mathrm{BAU}} > e$.

Second differentiating V_f, W_f, $V_f - V_c$ and $W_f - W_c$ with respect to Θ and using (4), (5), (A.13), (A.14) and (A.15) yields

$$\begin{aligned}
\frac{\mathrm{d}V_f}{\mathrm{d}\Theta} &= [1 + \alpha(n-m-1)]B'(e_f)\frac{\mathrm{d}e_f}{\mathrm{d}\Theta} + \alpha m B'(e_c)\frac{\mathrm{d}e_c}{\mathrm{d}\Theta} - [1 + \alpha(n-1)]D'\frac{\mathrm{d}e}{\mathrm{d}\Theta} \\
&= [1 + \alpha(n-1)]D'\left\{(1-\alpha)\frac{\mathrm{d}e_f}{\mathrm{d}\Theta} + \alpha\left[(n-m)\frac{\mathrm{d}e_f}{\mathrm{d}\Theta} + m\Theta\frac{\mathrm{d}e_c}{\mathrm{d}\Theta}\right] - \frac{\mathrm{d}e}{\mathrm{d}\Theta}\right\} \\
&= \frac{\alpha m[1 + \alpha(n-1)]^2\{(n-m)[1 + \alpha(n-1)]D'' - B''(e_f)\}(D')^2}{B''(e_c)B''(e_f) - [1 + \alpha(n-1)][(n-m)B''(e_c) + m\Theta B''(e_f)]D''}\left\{\frac{m}{1 + \alpha(m-1)}\right. \\
&\quad \left. + \frac{(1-\alpha)\{[1 - \alpha(m-1)(n-m-1)][1 + \alpha(n-1)]D'' - B''(e_f)\}}{\alpha[1 + \alpha(m-1)]\{(n-m)[1 + \alpha(n-1)]D'' - B''(e_f)\}} - \Theta\right\},
\end{aligned} \tag{A.16}$$

$$\frac{\mathrm{d}W_f}{\mathrm{d}\Theta} = B'(e_f)\frac{\mathrm{d}e_f}{\mathrm{d}\Theta} - D'\frac{\mathrm{d}e}{\mathrm{d}\Theta} > 0, \tag{A.17}$$

$$\frac{\mathrm{d}(V_f - V_c)}{\mathrm{d}\Theta} = (1-\alpha)\frac{\mathrm{d}(W_f - W_c)}{\mathrm{d}\Theta} = (1-\alpha)\left[B'(e_f)\frac{\mathrm{d}e_f}{\mathrm{d}\Theta} - B'(e_c)\frac{\mathrm{d}e_c}{\mathrm{d}\Theta}\right] > 0. \quad \text{(A.18)}$$

(A.16) yields $\frac{\mathrm{d}V_f}{\mathrm{d}\Theta} > 0$ for $\alpha \leq \frac{1}{(m-1)(n-m-1)}\left(\geq \frac{4}{(n-2)^2}\right)$ and $\Theta \leq \frac{m}{1+\alpha(m-1)}$, which implies $V_f > V_i^{\mathrm{BAU}}$ for $\alpha \leq \frac{4}{(n-2)^2}$. (A.17) implies $W_f > W_i^{\mathrm{BAU}}$. Finally, (A.18) implies $V_f > V_c$ and $W_f > W_c$.

Third suppose $e \leq e^{\mathrm{SO}}$. Then, the right-hand sides of (4) and (5) would be smaller than $nD'(e^{\mathrm{SO}})$, such that the left-hand sides would have to be smaller than $B'(e_i^{\mathrm{SO}})$, implying $e_c, e_f > e_i^{\mathrm{SO}}$ and contradicting $e \leq e^{\mathrm{SO}}$. Thus, $e > e^{\mathrm{SO}}$. □

Proof of Lemma 1

Totally differentiating (4), (5) and $e = me_c + (n-m)e_f$ yields

$$B''(e_f)\mathrm{d}e_f = [1+\alpha(n-1)]D''\mathrm{d}e + (n-1)D'\mathrm{d}\alpha, \quad \text{(A.19)}$$

$$B''(e_c)\mathrm{d}e_c = \frac{1+\alpha(n-1)}{1+\alpha(m-1)}mD''\mathrm{d}e + \frac{(1-\alpha)[1+\alpha(n-1)]}{[1+\alpha(m-1)]^2}D'\mathrm{d}m + \frac{n-m}{[1+\alpha(m-1)]^2}mD'\mathrm{d}\alpha, \quad \text{(A.20)}$$

$$\mathrm{d}e = m\mathrm{d}e_c + (n-m)\mathrm{d}e_f + (e_c - e_f)\mathrm{d}m. \quad \text{(A.21)}$$

Solving for $\mathrm{d}e_f$, $\mathrm{d}e_c$ and $\mathrm{d}e$ yields

$$\begin{aligned}\Lambda\mathrm{d}e_f &= \{(1-\alpha)[1+\alpha(n-1)]mD' - (e_f - e_c)[1+\alpha(m-1)]^2B''(e_c)\}[1+\alpha(n-1)]D''\mathrm{d}m \\ &\quad - \{m^2(m-1)[1+\alpha(n-1)]^2D'' - (n-1)[1+\alpha(m-1)]^2B''(e_c)\}D'\mathrm{d}\alpha,\end{aligned} \quad \text{(A.22)}$$

$$\begin{aligned}\Lambda\mathrm{d}e_c &= -\{(1-\alpha)\{[1+\alpha(n-1)](n-m)D'' - B''(e_f)\}D' + (e_f - e_c)[1+\alpha(m-1)]mD'' \\ &\quad \cdot B''(e_f)\}[1+\alpha(n-1)]\mathrm{d}m - m(n-m)\{[1+\alpha(n-1)]^2(m-1)D'' + B''(e_f)\}D'\mathrm{d}\alpha,\end{aligned} \quad \text{(A.23)}$$

$$\begin{aligned}\Lambda\mathrm{d}e &= \{(1-\alpha)[1+\alpha(n-1)]mD' - (e_f - e_c)[1+\alpha(m-1)]^2B''(e_c)\}B''(e_f)\mathrm{d}m \\ &\quad + (n-m)\{[1+\alpha(m-1)]^2(n-1)B''(e_c) + m^2B''(e_f)\}D'\mathrm{d}\alpha,\end{aligned} \quad \text{(A.24)}$$

where

$$\begin{aligned}\Lambda &:= -[1+\alpha(m-1)]\{[1+\alpha(m-1)]\{(n-m)[1+\alpha(n-1)]D'' - B''(e_f)\}B''(e_c) \\ &\quad + m^2[1+\alpha(n-1)]D''B''(e_f)\} > 0.\end{aligned}$$

First (A.22) [(A.24)] yields $\frac{\mathrm{d}e_f}{\mathrm{d}m} > 0$ and $\frac{\mathrm{d}e_f}{\mathrm{d}\alpha} < 0$ [$\frac{\mathrm{d}e}{\mathrm{d}m} < 0$ and $\frac{\mathrm{d}e}{\mathrm{d}\alpha} < 0$].

Second differentiating V_f and W_f with respect to m and using (4), (5), (A.22), (A.23) and (A.24) yields

$$\frac{\mathrm{d}V_f}{\mathrm{d}m} = [1+\alpha(n-1)]D'\left\{(1-\alpha)\frac{\mathrm{d}e_f}{\mathrm{d}m} + \alpha\left[(n-m)\frac{\mathrm{d}e_f}{\mathrm{d}m} + \frac{m^2}{1+\alpha(m-1)}\frac{\mathrm{d}e_c}{\mathrm{d}m}\right] - \frac{\mathrm{d}e}{\mathrm{d}m}\right\}$$
$$= \frac{[1+\alpha(n-1)]D'}{[1+\alpha(m-1)]\Lambda}\{(1-\alpha)[1+\alpha(n-1)]2\{[1-(m-1)(n-m-1)\alpha][1+\alpha(n-1)]D''$$
$$- B''(e_f)\}mD' - (e_f-e_c)[1+\alpha(m-1)]\{[1+\alpha(n-m-1)][1+\alpha(m-1)]^2[1+\alpha(n$$
$$- 1)]D''B''(e_c) - [1+\alpha(m-1)]^2B''(e_f)B''(e_c) + \alpha[1+\alpha(n-1)]m^3D''B''(e_f)\}\}, \tag{A.25}$$

$$\frac{\mathrm{d}W_f}{\mathrm{d}m} = D'\left\{[1+\alpha(n-1)]\frac{\mathrm{d}e_f}{\mathrm{d}m} - \frac{\mathrm{d}e}{\mathrm{d}m}\right\} > 0, \tag{A.26}$$

such that $\frac{\mathrm{d}V_f}{\mathrm{d}m} > 0$ for $\alpha \leq \frac{1}{(m-1)(n-m-1)}\left(\geq \frac{4}{(n-2)^2}\right)$.

Third differentiating W with respect to α and using (4), (5) and (A.21) yields

$$\begin{aligned}\frac{\mathrm{d}W}{\mathrm{d}\alpha} &= D'\left\{[1+\alpha(n-1)]\left[(n-m)\frac{\mathrm{d}e_f}{\mathrm{d}\alpha} + \frac{m^2}{1+\alpha(m-1)}\frac{\mathrm{d}e_c}{\mathrm{d}\alpha}\right] - n\frac{\mathrm{d}e}{\mathrm{d}\alpha}\right\} \\ &= -\frac{(1-\alpha)(n-m)}{1+\alpha(m-1)}D'\left\{(m-1)[1+\alpha(n-1)]\frac{\mathrm{d}e_f}{\mathrm{d}\alpha} + \frac{\mathrm{d}e}{\mathrm{d}\alpha}\right\} > 0.\end{aligned} \tag{A.27}$$

Finally, differentiating (6) and (7) with respect to α yields

$$\begin{aligned}\frac{\mathrm{d}R'_f}{\mathrm{d}\alpha} &= \frac{(n-m)(n-1)D''B''(e_f)}{\{(n-m)[1+\alpha(n-1)]D'' - B''(e_f)\}^2} \\ &\quad + \frac{(n-m)[1+\alpha(n-1)]D'''B''(e_f)}{\{(n-m)[1+\alpha(n-1)]D'' - B''(e_f)\}^2}\frac{\mathrm{d}e}{\mathrm{d}\alpha} \\ &\quad - \frac{(n-m)[1+\alpha(n-1)]D''B'''(e_f)}{\{(n-m)[1+\alpha(n-1)]D'' - B''(e_f)\}^2}\frac{\mathrm{d}e_f}{\mathrm{d}\alpha},\end{aligned} \tag{A.28}$$

$$\begin{aligned}\frac{\mathrm{d}R'_c}{\mathrm{d}\alpha} &= \frac{m^2(n-m)D''B''(e_c)}{\{m^2[1+\alpha(n-1)]D''(e) - [1+\alpha(m-1)]B''(e_c)\}^2} \\ &\quad + \frac{m^2[1+\alpha(n-1)][1+\alpha(m-1)]D'''B''(e_c)}{\{m^2[1+\alpha(n-1)]D''(e) - [1+\alpha(m-1)]B''(e_c)\}^2}\frac{\mathrm{d}e}{\mathrm{d}\alpha} \\ &\quad - \frac{m^2[1+\alpha(n-1)][1+\alpha(m-1)]D''B'''(e_c)}{\{m^2[1+\alpha(n-1)]D''(e) - [1+\alpha(m-1)]B''(e_c)\}^2}\frac{\mathrm{d}e_c}{\mathrm{d}\alpha},\end{aligned} \tag{A.29}$$

such that $\frac{\mathrm{d}e}{\mathrm{d}\alpha} < 0$ implies $\frac{\mathrm{d}R'_f}{\mathrm{d}\alpha}, \frac{\mathrm{d}R'_c}{\mathrm{d}\alpha} < 0$ for $D''' \leq 0$ and $B''' = 0$. □

Proof of Proposition 2

$$a - be_f = [1+\alpha(n-1)]de, \tag{A.30}$$

$$a - be_c = \frac{1+\alpha(n-1)}{1+\alpha(m-1)} mde, \tag{A.31}$$

$$e = (n-m)e_f + me_c. \tag{A.32}$$

Solving for e_f, e_c and e yields

$$e_f = \frac{1+\alpha(m-1)+m(m-1)(1-\alpha)[1+\alpha(n-1)]\frac{d}{b}}{\Omega}\frac{a}{b} > 0, \tag{A.33}$$

$$e_c = \frac{1+\alpha(m-1)-(n-m)(m-1)(1-\alpha)[1+\alpha(n-1)]\frac{d}{b}}{\Omega}\frac{a}{b}, \tag{A.34}$$

$$e = \frac{1+\alpha(m-1)}{\Omega}\frac{na}{b} > 0, \tag{A.35}$$

where

$$\Omega := 1+\alpha(m-1)+[1+\alpha(n-1)]\{[1+\alpha(m-1)]n+(1-\alpha)m(m-1)\}\frac{d}{b} > 0.$$

Note that $\frac{\partial[}{\partial 2}]\frac{e_c\Omega}{1+\alpha(m-1)}m > 0$. For $\alpha = 0$, $\frac{e_c\Omega}{1+\alpha(m-1)}$ is minimal at $m = \frac{n+1}{2}$, and then $\frac{e_c\Omega}{1+\alpha(m-1)}$ is non-negative if and only if $d \le \bar{d} := \frac{4b}{(n-1)^2}$, which is thus an upper bound for d. Using this upper bound in (A.34) yields

$$e_c = \frac{a}{4b^2\Omega}\Big\{\big\{\alpha(m-1)[(n-1)(n-m)(4\alpha-3)+(m+3)(n-m)+(m-1)^2] \\ +(n+1-2m)^2\big\}d+(n-1)^2[1+\alpha(m-1)](\bar{d}-d)\Big\} > 0 \Longleftarrow d \le \bar{d} \wedge \alpha \ge 3/4.$$

For $\alpha > 0$, $\frac{e_c\Omega}{1+\alpha(m-1)}$ is minimal at

$$m = 1+\frac{\sqrt{1+\alpha(n-1)}-1}{\alpha} = \frac{n+1}{2} - \frac{\alpha(n-1)^2/2}{2\sqrt{1+\alpha(n-1)}+2+\alpha(n-1)} \in \left(1, \frac{n+1}{2}\right),$$

and then $\frac{e_c\Omega}{1+\alpha(m-1)}$ is non-negative if and only if

$$d \le \frac{\alpha^2 b}{(1-\alpha)\{1+\alpha(n-1)-\sqrt{1+\alpha(n-1)}\}^2}.$$

Using (A.33), (A.34) and (A.35) yields

$$
\begin{aligned}
V_f &= [1+\alpha(n-m-1)]\Big[ae_f - \frac{b}{2}e_f^2 - \frac{d}{2}e^2\Big] + \alpha m\Big[ae_c - \frac{b}{2}e_c^2 - \frac{d}{2}e^2\Big] \\
&= \frac{a^2}{2b\Omega^2}[1+\alpha(n-1)]\{[1+\alpha(m-1)]^2 - [1+\alpha(m-1)]\{[1+\alpha(m-1)][n^2-2m^2 \\
&\quad - 2n + 2m - 2\alpha(n-1)(n-m)] - 2\alpha m^2(n-m)\}\frac{d}{b} + (1-\alpha)^2[1+\alpha(n-1)]m(m \\
&\quad - 1)\{m^2 + 2n - m - \alpha(m-1)(n^2 - nm - 2n + m)\}\Big(\frac{d}{b}\Big)^2\},
\end{aligned}
\tag{A.36}
$$

$$
\begin{aligned}
V_c &= [1+\alpha(n-m)]\Big[ae_f - \frac{b}{2}e_f^2 - \frac{d}{2}e^2\Big] + [1+\alpha(m-1)]\Big[ae_c - \frac{b}{2}e_c^2 - \frac{d}{2}e^2\Big] \\
&= V_f - \frac{a^2}{2b\Omega^2}n^2(m-1)(1-\alpha)^2[1+\alpha(n-1)]^2[m+1+\alpha(m-1)]\Big(\frac{d}{b}\Big)^2.
\end{aligned}
\tag{A.37}
$$

The internal stability condition reads

$$
V_c(m) - V_f(m-1) = \frac{a^2n^2(m-1)(1-\alpha)^2[1+\alpha(n-1)]^2\Big(\frac{d}{b}\Big)^2}{2b\Omega(m)^2\Omega(m-1)^2}\Phi(m), \tag{A.38}
$$

where

$$
\begin{aligned}
\Phi(m) &:= -[1+\alpha(m-1)][m-3+\alpha(m-2)^2] - [1+\alpha(n-1)]\{2[(n-m)(m-1) + (m \\
&\quad - 3)^3 + 6(m-3)^2 + 11(m-3) + 4] + 2\alpha\{(n-m)[2(m-2)^2 + 5(m-2) + 1] \\
&\quad + (m-2)[(m-2)^3 + 4(m-2)^2 + 6(m-2) + 2]\} + \alpha^2(m-2)[2(n-m)(m^2 \\
&\quad + m - 3) + (m-1)(m-2)] + 2\alpha^3(n-m)(m-1)(m-2)^2\}\frac{d}{b} - [1+\alpha(n-1)]^2 \\
&\quad \cdot \{[n(m+1) + m(m-1)^2][n + m(m-3)] + \alpha\{(n-m)^2[2(m-2)^2 + 10(m-2) \\
&\quad + 3] + 2(n-m)[(m-2)^4 + 7(m-2)^3 + 16(m-2)^2 + 14(m-2) + 2] + m^2(m \\
&\quad - 2)^2\} + \alpha^2(n-m)(m-2)\{(n-m)[(m-2)^2 + 9(m-2) + 6] + 2(m-2)^3 + 8 \\
&\quad \cdot (m-2)^2 + 12(m-2) + 4\} + \alpha^3(n-m)(2n-2m+1)(m-1)(m-2)^2\}\Big(\frac{d}{b}\Big)^2 \\
&\quad < 0 \Longleftarrow m \geq 3,
\end{aligned}
\tag{A.39}
$$

such that all coalitions $m \geq 3$ are internally unstable, which proves the first bullet of the proposition. Furthermore,

$$\begin{aligned}\Phi(2)|_{\alpha=0} &= 1 - 2(n-4)\frac{d}{b} - (n-2)(3n+2)\left(\frac{d}{b}\right)^2 \gtreqless 0 \\ &\iff \frac{d}{b} \lesseqgtr \frac{2\sqrt{n^2-3n+3} - n + 4}{3n^2 - 4n - 4},\end{aligned} \tag{A.40}$$

$$\begin{aligned}&= \frac{(n-1)^4}{16b^2}\left\{\frac{(n-12)^4 + 36(n-12)^3 + 438(n-12)^2 + 1860(n-12) + 817}{(n-1)^4}d^2\right. \\ &\left. + \frac{2(n-3)^2+16}{(n-1)^2}(\bar{d}-d)d + (\bar{d}-d)^2\right\} > 0 \Longleftarrow n \geq 12 \wedge d \leq \bar{d},\end{aligned} \tag{A.41}$$

$$\begin{aligned}&= -\frac{(n-1)^4}{16b^2}\left\{\frac{\left(\frac{11-n}{n-3}\right)^4 + 16\left(\frac{11-n}{n-3}\right)^3 + 66\left(\frac{11-n}{n-3}\right)^2 + 88\left(\frac{11-n}{n-3}\right) + 5}{32\left(\frac{n-1}{n-3}\right)^4}d^2\right. \\ &\left. - \frac{2(n-3)^2+16}{(n-1)^2}(\bar{d}-d)d - (\bar{d}-d)^2\right\} < 0 \Longleftarrow n \leq 11 \wedge d = \bar{d},\end{aligned} \tag{A.42}$$

which proves the second bullet of the proposition. Finally,

$$\frac{\Phi(2)}{1+\alpha} = 1 - \frac{2[n-4+\alpha(n-2)]}{(1+\alpha)/[1+\alpha(n-1)]}\frac{d}{b} - \frac{(n-2)[3n+2+\alpha(3n-2)]}{(1+\alpha)/[1+\alpha(n-1)]^2}\left(\frac{d}{b}\right)^2, \tag{A.43}$$

where

$$\frac{\partial\left(\frac{\Phi(2)}{1+\alpha}\right)}{\partial\left(\frac{d}{b}\right)} < 0 \Longleftarrow n \geq 4, \tag{A.44}$$

$$\begin{aligned}\frac{\partial\left(\frac{\Phi(2)}{1+\alpha}\right)}{\partial\alpha} &= -\frac{2(n-2)[(n+1)(3n-4) + (2+\alpha)\alpha(n-1)(3n-2)]}{(1+\alpha)^2/[1+\alpha(n-1)]}\left(\frac{d}{b}\right)^2 \\ &- \frac{2(n-2)[n-3+(2+\alpha)\alpha(n-1)]}{(1+\alpha)^2}\frac{d}{b} < 0 \Longleftarrow n \geq 3,\end{aligned} \tag{A.45}$$

$$\begin{aligned}\frac{\partial\left(\frac{\Phi(2)}{1+\alpha}\right)}{\partial n} &= -\frac{2[3n-2+6\alpha(n^2-n-1)+\alpha^2(6n^2-15n+8)]}{(1+\alpha)/[1+\alpha(n-1)]}\left(\frac{d}{b}\right)^2 \\ &- \frac{2[1+2\alpha(n-2)+\alpha^2(2n-3)]}{(1+\alpha)}\frac{d}{b} < 0 \Longleftarrow n \geq 2,\end{aligned} \tag{A.46}$$

such that $m = 2$ is internally stable if $\frac{d}{b}$, α and n are sufficiently small. Since all coalitions $m \geq 3$ are internally unstable from (A.39), and the condition for $m = 2$ to be stable

becomes stricter as α increases from (A.45), the coalition size weakly decreases with α, which proves the third bullet of the proposition. □

Effects of Altruism on the Internal Stability Condition

From (11), the direct effect of altruism on the internal stability condition reads

$$\begin{aligned}&\frac{\partial\left[V_c(m)-V_f(m-1)\right]}{\partial\alpha}\\&=\left[(m-1)W_c(m)+(n-m)W_f(m)\right]-\left[(m-1)W_c(m-1)+(n-m)W_f(m-1)\right],\end{aligned}\tag{A.47}$$

which is positive if $(m-1)\frac{\mathrm{d}W_c(m)}{\mathrm{d}m}+(n-m)\frac{\mathrm{d}W_f(m)}{\mathrm{d}m}>0$. Using (4), (5) and (A.21) yields

$$\begin{aligned}&(m-1)\frac{\mathrm{d}W_c(m)}{\mathrm{d}m}+(n-m)\frac{\mathrm{d}W_f(m)}{\mathrm{d}m}\\&=(m-1)\left[B'(e_c)\frac{\mathrm{d}e_c}{\mathrm{d}m}-D'\frac{\mathrm{d}e}{\mathrm{d}m}\right]+(n-m)\left[B'(e_f)\frac{\mathrm{d}e_f}{\mathrm{d}m}-D'\frac{\mathrm{d}e}{\mathrm{d}m}\right]\\&=D'\left\{(m-1)\left[\frac{1+\alpha(n-1)}{1+\alpha(m-1)}m\frac{\mathrm{d}e_c}{\mathrm{d}m}-\frac{\mathrm{d}e}{\mathrm{d}m}\right]+(n-m)\left[[1+\alpha(n-1)]\frac{\mathrm{d}e_f}{\mathrm{d}m}-\frac{\mathrm{d}e}{\mathrm{d}m}\right]\right\}\\&=\frac{(m-1)(e_f-e_c)+(n-m)\left\{[2-m+\alpha(m-1)]\frac{\mathrm{d}e_f}{\mathrm{d}m}-[1+\alpha(n-1)]^{-1}\frac{\mathrm{d}e}{\mathrm{d}m}\right\}}{[1+\alpha(n-1)]^{-1}[1+\alpha(m-1)]/D'},\end{aligned}\tag{A.48}$$

such that $\frac{\partial\left[V_c(m)-V_f(m-1)\right]}{\partial\alpha}>0$ for $m\in\{2,n\}$. Using specification (9), it can be shown that $\frac{\partial\left[V_c(m)-V_f(m-1)\right]}{\partial\alpha}>0$ for $m\in[2,n]$. The corresponding Maple file is available on request.

From (12), the indirect effects of altruism on the internal stability condition read

$$(1-\alpha)\frac{\mathrm{d}W_c(m)}{\mathrm{d}\alpha}+\alpha\frac{\mathrm{d}W(m)}{\mathrm{d}\alpha}-\left[(1-\alpha)\frac{\mathrm{d}W_f(m-1)}{\mathrm{d}\alpha}+\alpha\frac{\mathrm{d}W(m-1)}{\mathrm{d}\alpha}\right].\tag{A.49}$$

Using (4), (5) and (A.21) yields

$$\begin{aligned}&(1-\alpha)\frac{\mathrm{d}W_c(m)}{\mathrm{d}\alpha}+\alpha\frac{\mathrm{d}W(m)}{\mathrm{d}\alpha}\\&=(1-\alpha)\left[B'(e_c(m))\frac{\mathrm{d}e_c(m)}{\mathrm{d}\alpha}-D'\frac{\mathrm{d}e(m)}{\mathrm{d}\alpha}\right]\\&\quad+\alpha\left[mB'(e_c(m))\frac{\mathrm{d}e_c(m)}{\mathrm{d}\alpha}+(n-m)B'(e_f(m))\frac{\mathrm{d}e_f(m)}{\mathrm{d}\alpha}-nD'\frac{\mathrm{d}e(m)}{\mathrm{d}\alpha}\right]\\&=[1+\alpha(n-1)]D'\left[m\frac{\mathrm{d}e_c(m)}{\mathrm{d}\alpha}+\alpha(n-m)\frac{\mathrm{d}e_f(m)}{\mathrm{d}\alpha}-\frac{\mathrm{d}e(m)}{\mathrm{d}\alpha}\right]\\&=-(1-\alpha)[1+\alpha(n-1)](n-m)D'\frac{\mathrm{d}e_f(m)}{\mathrm{d}\alpha}>0\end{aligned}\tag{A.50}$$

and

$$
\begin{aligned}
&(1-\alpha)\frac{\mathrm{d}W_f(m-1)}{\mathrm{d}\alpha}+\alpha\frac{\mathrm{d}W(m-1)}{\mathrm{d}\alpha}\\
&=(1-\alpha)\left[B'(e_f(m-1))\frac{\mathrm{d}e_f(m-1)}{\mathrm{d}\alpha}-D'\frac{\mathrm{d}e(m-1)}{\mathrm{d}\alpha}\right]\\
&\quad+\alpha\left[(m-1)B'(e_c(m-1))\frac{\mathrm{d}e_c(m-1)}{\mathrm{d}\alpha}+(n-m+1)B'(e_f(m-1))\frac{\mathrm{d}e_f(m-1)}{\mathrm{d}\alpha}-nD'\frac{\mathrm{d}e(m-1)}{\mathrm{d}\alpha}\right]\\
&=[1+\alpha(n-1)]D'\left\{[1+\alpha(n-m)]\frac{\mathrm{d}e_f(m-1)}{\mathrm{d}\alpha}+\frac{\alpha(m-1)^2}{1+\alpha(m-2)}\frac{\mathrm{d}e_c(m-1)}{\mathrm{d}\alpha}-\frac{\mathrm{d}e(m-1)}{\mathrm{d}\alpha}\right\}\\
&=-\frac{(1-\alpha)[1+\alpha(n-1)]D'}{1+\alpha(m-2)}\left\{[\alpha(m-2)(n-m)-1]\frac{\mathrm{d}e_f(m-1)}{\mathrm{d}\alpha}+\frac{\mathrm{d}e(m-1)}{\mathrm{d}\alpha}\right\},
\end{aligned}
\tag{A.51}
$$

such that $(1-\alpha)\frac{\mathrm{d}W_c(m)}{\mathrm{d}\alpha}+\alpha\frac{\mathrm{d}W(m)}{\mathrm{d}\alpha}>0$ and $(1-\alpha)\frac{\mathrm{d}W_f(2-1)}{\mathrm{d}\alpha}+\alpha\frac{\mathrm{d}W(2-1)}{\mathrm{d}\alpha}=[1+\alpha(n-1)]\frac{\mathrm{d}W_i^{\mathrm{BAU}}}{\mathrm{d}\alpha}>0$ from Appendix A.1. Using specification (9), it can be shown that $(1-\alpha)\frac{\mathrm{d}W_f(m-1)}{\mathrm{d}\alpha}+\alpha\frac{\mathrm{d}W(m-1)}{\mathrm{d}\alpha}>0$ for $m\in[2,n-2]$. The corresponding Maple file is available on request.

Proof of Proposition 3

Without altruistic preferences at the second stage of the game, the emissions for a given coalition size are given by substituting $\alpha=0$ into (A.33), (A.34) and (A.35), and the material welfare levels for a given coalition size are given by substituting $\alpha=0$ into (A.36) and (A.37). Using these results, the internal stability condition reads

$$
\begin{aligned}
V_c(m)-V_f(m-1)&=(1-\alpha)W_c(m)+\alpha W(m)-\left[(1-\alpha)W_f(m-1)+\alpha W(m-1)\right]\\
&=\frac{(1+2\alpha)n^2(m-1)\frac{a^2}{2b}\left(\frac{d}{b}\right)^2}{\left[1+(m^2+n-3m+2)\frac{d}{b}\right]^2\left[1+(m^2+n-m)\frac{d}{b}\right]^2}\varphi(m),
\end{aligned}
\tag{A.52}
$$

where

$$
\begin{aligned}
\varphi(m):=\frac{\alpha}{1+2\alpha}&\left\{4N_3+3+2[2N_3^2+N_3(2M_2^2+4M_2+13)+2M_2^3+9M_2^2+7M_2+9]\frac{d}{b}\right.\\
&+[7N_3^2+N_3[4M_2^3+22M_2^2+22M_2+34]+4M_2^5+23M_2^4+54M_2^3+85M_2^2\\
&\left.+50M_2+27]\left(\frac{d}{b}\right)^2\right\}-\left\{m-3+2[(n-m)(m-1)+(m-3)(m^2+2)+4]\frac{d}{b}\right.\\
&\left.+[n-m+m(m-2)][(n-m)(m+1)+m(m^2-m+2)]\left(\frac{d}{b}\right)^2\right\},
\end{aligned}
\tag{A.53}
$$

where $N_i:=n-i$ and $M_i:=m-i$. From (A.39) and (A.53), $\Phi(m)|_{\alpha=0}=\varphi(m)|_{\alpha=0}$, and from the proof of Proposition 2, $\Phi(2)|_{\alpha=0}>0$ for $n\geq 12$, which proves the second bullet of

the proposition. From (A.53), $\varphi(m)$ increases with $\frac{\alpha}{1+2\alpha}$ and, thus, with α, which proves the third bullet of the proposition. Furthermore,

$$\begin{aligned}\varphi(n) = \frac{(n-1)^4}{16b^2}\Bigg\{ & \frac{50N_2^5 + 305N_2^4 + 720N_2^3 + 790N_2^2 + 358N_2 + 17}{(1+2\alpha)(n-1)^4} \\ & \cdot \left[\alpha - \frac{4}{7} + \frac{N_3(5N_3^2 + 24N_3 + 8)}{7(10N_3^3 + 55N_3^2 + 100N_3 + 56)}\right] d^2 \\ & + \frac{2(10N_2^3 + 33N_2^2 + 36N_2 + 9)}{(1+2\alpha)(n-1)^2} \\ & \cdot \left[\alpha - \frac{4}{7} + \frac{5n^3 + 11n^2 - 9n + 33}{7(10N_2^3 + 33N_2^2 + 36N_2 + 9)}\right](\bar{d} - d)d \\ & + \frac{2N_2 + 1}{1+2\alpha}\left[\alpha - \frac{4}{7} + \frac{n+9}{7(2N_2+1)}\right](\bar{d} - d)^2\Bigg\} > 0 \Longleftarrow \alpha \geq \frac{4}{7} \wedge d \leq \bar{d},\end{aligned} \tag{A.54}$$

$$\begin{aligned}= & -\frac{n^2(n-2)(2n^2 - 3n + 2)}{1+2\alpha}\left[\frac{3}{7} - \alpha + \frac{n^2 + 2n + 8}{7(2n^2 - 3n + 2)}\right]\left(\frac{d}{b}\right)^2 \\ & - \frac{2(2N_2^3 + 7N_2^2 + 8N_2 + 2)}{1+2\alpha}\left[\frac{3}{7} - \alpha + \frac{N_6^3 + 12N_6^2 + 38N_6 + 4}{7(2N_2^3 + 7N_2^2 + 8N_2 + 2)}\right]\frac{d}{b} \\ & - \frac{2N_2 + 1}{1+2\alpha}\left[\frac{3}{7} - \alpha + \frac{N_{12}}{7(2N_2+1)}\right] < 0 \Longleftarrow \alpha \leq \frac{3}{7} \wedge n \geq 12,\end{aligned} \tag{A.55}$$

which proves the third bullet of the proposition. Furthermore,

$$\begin{aligned}\frac{\partial\varphi(m)}{\partial m} = & \Bigg\{1 + 2(n - m + 3m^2 - 7m + 3)\frac{d}{b} + [(n-m)^2 + (6m^2 - 6m - 2)(n-m) + 5m^4 \\ & - 14m^3 + 14m^2 - 8m]\left(\frac{d}{b}\right)^2\Bigg\} \Bigg/ \Bigg\{m - 3 + 2[(m-1)(n-m) + (m-3)(m^2+2) \\ & + 4]\frac{d}{b} + [n - m + m(m-2)][(m+1)(n-m) + m(m^2 - m + 2)]\left(\frac{d}{b}\right)^2\Bigg\}\varphi(m) - \Psi,\end{aligned} \tag{A.56}$$

where $\Psi > 0$ for $n \geq 4$ and $m \geq 3$. The corresponding Maple file is available on request. $\varphi(\underline{m}) \leq 0$ for some $\underline{m} \geq 3$ implies $\frac{\partial\varphi(\underline{m})}{\partial m} < 0$ and, thus, $\varphi(\bar{m}) < 0$ for all $\bar{m} \geq \underline{m}$. Furthermore, $\varphi(m) \geq 0 \Longleftrightarrow V_c(m) - V_f(m-1) \geq 0 \Longleftrightarrow V_f(m-1) - V_c(m) \leq 0$. Thus, an internally stable coalition m implies an externally unstable coalition $m-1$. Consequently, there is at most one internally and externally stable coalition, which proves the first bullet of the proposition. Finally, note that

$$W = \frac{\{b^2 - b[n(n-2) - 2m(m-1)]d - m(m-1)^2(n-m)d^2\}na^2}{2b[b + (m^2 + n - m)d]^2}, \tag{A.57}$$

$$\frac{\partial W}{\partial m} = \frac{[(4m-2)(n-m)+(m-1)^2]n^2ad^2}{2[b+(m^2+n-m)d]^2}e_c + \frac{(m-1)^4n^2(n-2)a^2d^3}{2(n-1)^2b[b+(m^2+n-m)d]^3} \cdot \left[2\left(\frac{n-m}{m-1}\right)^3 + (4n-1)\left(\frac{n-m}{m-1}\right)^2 + (n-1)\left(\frac{n-m}{m-1}\right) + \frac{n^2}{n-2}\right] > 0, \tag{A.58}$$

such that $[1+\alpha(n-1)]\frac{\partial W}{\partial m} = \frac{\partial V}{\partial m} > 0$. □

Stackelberg Game

Derivation of (13)

The first-order condition of (3) reads

$$[1+\alpha(m-1)]\left\{B'(e_c) - \frac{m[1+\alpha(n-1)]}{1+\alpha(m-1)}D'\left[1+\frac{\mathrm{d}(n-m)e_f}{\mathrm{d}e_c}\right] + \frac{\alpha m(n-m)}{1+\alpha(m-1)}B'(e_f)\frac{\mathrm{d}e_f}{\mathrm{d}e_c}\right\} = 0. \tag{A.59}$$

Substituting (4) and rearranging yields (13). The second-order condition of (3) reads

$$[1+\alpha(m-1)]\left\{B''(e_c) - \frac{m[1+\alpha(n-1)]}{1+\alpha(m-1)}D''\left[1+\frac{\mathrm{d}(n-m)e_f}{\mathrm{d}e_c}\right]^2 + \frac{\alpha m(n-m)}{1+\alpha(m-1)}B''(e_f)\left(\frac{\mathrm{d}e_f}{\mathrm{d}e_c}\right)^2 - \frac{m}{1+\alpha(m-1)}\{[1+\alpha(n-1)]D' - \alpha B'(e_f)\}\frac{(n-m)[1+\alpha(n-1)]}{\{(n-m)[1+\alpha(n-1)]D'' - B''(e_f)\}^2} \cdot \left[B''(e_f)D''\left[1+\frac{\mathrm{d}(n-m)e_f}{\mathrm{d}e_c}\right] - D''B'''(e_f)\frac{\mathrm{d}e_f}{\mathrm{d}e_c}\right]\right\} < 0, \tag{A.60}$$

which is fulfilled if $D''' \leq 0, B''' \geq 0$.

Proof of Proposition 4

From (4), (6) and (13), the equilibrium is characterised by

$$B'(e_f) = [1+\alpha(n-1)]D', \tag{A.61}$$

$$B'(e_c) = [1+\alpha(n-1)]\theta D', \tag{A.62}$$

$$e = me_c + (n-m)e_f. \tag{A.63}$$

First differentiating (A.61), (A.62) and (A.63) with respect to θ yields

$$B''(e_f)\frac{\mathrm{d}e_f}{\mathrm{d}\theta} = [1+\alpha(n-1)]D''\frac{\mathrm{d}e}{\mathrm{d}\theta}, \tag{A.64}$$

$$B''(e_c)\frac{\mathrm{d}e_c}{\mathrm{d}\theta} = [1+\alpha(n-1)]\left[\theta D''\frac{\mathrm{d}e}{\mathrm{d}\theta} + D'\right], \tag{A.65}$$

$$\frac{\mathrm{d}e}{\mathrm{d}\theta} = m\frac{\mathrm{d}e_c}{\mathrm{d}\theta} + (n-m)\frac{\mathrm{d}e_f}{\mathrm{d}\theta}. \tag{A.66}$$

Solving for $\frac{\mathrm{d}e}{\mathrm{d}\theta}$, $\frac{\mathrm{d}e_f}{\mathrm{d}\theta}$ and $\frac{\mathrm{d}e_c}{\mathrm{d}\theta}$ yields

$$\frac{\mathrm{d}e_f}{\mathrm{d}\theta} = \frac{m[1+\alpha(n-1)]^2 D''D'}{B''(e_c)B''(e_f) - [1+\alpha(n-1)][(n-m)B''(e_c) + m\theta B''(e_f)]D''} > 0, \tag{A.67}$$

$$\frac{\mathrm{d}e_c}{\mathrm{d}\theta} = -\frac{[1+\alpha(n-1)]\{(n-m)[1+\alpha(n-1)]D'' - B''(e_f)\}D'}{B''(e_c)B''(e_f) - [1+\alpha(n-1)][(n-m)B''(e_c) + m\theta B''(e_f)]D''} < 0, \tag{A.68}$$

$$\frac{\mathrm{d}e}{\mathrm{d}\theta} = \frac{m[1+\alpha(n-1)]B''(e_f)D'}{B''(e_c)B''(e_f) - [1+\alpha(n-1)][(n-m)B''(e_c) + m\theta B''(e_f)]D''} < 0. \tag{A.69}$$

Note that $\theta = 1 \iff e_f = e_c = e_i^{\mathrm{BAU}}$. Thus, $\theta \gtreqless 1 \iff e_f \gtreqless e_i^{\mathrm{BAU}} \gtreqless e_c \wedge e^{\mathrm{BAU}} \gtreqless e$.

Second differentiating V_i and W_i with respect to θ and using (4), (6), (13), (A.67), (A.68) and (A.69) yields

$$\begin{aligned}
\frac{\mathrm{d}V_f}{\mathrm{d}\theta} &= [1+\alpha(n-m-1)]B'(e_f)\frac{\mathrm{d}e_f}{\mathrm{d}\theta} + \alpha m B'(e_c)\frac{\mathrm{d}e_c}{\mathrm{d}\theta} - [1+\alpha(n-1)]D'\frac{\mathrm{d}e}{\mathrm{d}\theta} \\
&= [1+\alpha(n-1)]D'\left\{(1-\alpha)\frac{\mathrm{d}e_f}{\mathrm{d}\theta} + \alpha\left[(n-m)\frac{\mathrm{d}e_f}{\mathrm{d}\theta} + m\theta\frac{\mathrm{d}e_c}{\mathrm{d}\theta}\right] - \frac{\mathrm{d}e}{\mathrm{d}\theta}\right\} \\
&= \frac{\alpha m[1+\alpha(n-1)]^2\{(n-m)[1+\alpha(n-1)]D'' - B''(e_f)\}(D')^2}{B''(e_c)B''(e_f) - [1+\alpha(n-1)][(n-m)B''(e_c) + m\theta B''(e_f)]D''} \\
&\quad \cdot \left\{\frac{(1-\alpha)\{[1+\alpha(n-1)]D'' - B''(e_f)\}}{\alpha\{(n-m)[1+\alpha(n-1)]D'' - B''(e_f)\}} + 1 - \theta\right\}
\end{aligned} \tag{A.70}$$

$$\begin{aligned}
&= \frac{\alpha m[1+\alpha(n-1)]^2\{(n-m)[1+\alpha(n-1)]D'' - B''(e_f)\}(D')^2}{B''(e_c)B''(e_f) - [1+\alpha(n-1)][(n-m)B''(e_c) + m\theta B''(e_f)]D''} \\
&\quad \cdot \left\{\frac{(1-\alpha)\{[1+\alpha(n-1)]^2 D'' - B''(e_f)\}}{\alpha[1+\alpha(m-1)]\{(n-m)[1+\alpha(n-1)]D'' - B''(e_f)\}} + \tilde{\theta} - \theta\right\},
\end{aligned} \tag{A.71}$$

$$\begin{aligned}
\frac{\mathrm{d}V_c}{\mathrm{d}\theta} &= \alpha(n-m)B'(e_f)\frac{\mathrm{d}e_f}{\mathrm{d}\theta} + [1+\alpha(m-1)]B'(e_c)\frac{\mathrm{d}e_c}{\mathrm{d}\theta} - [1+\alpha(n-1)]D'\frac{\mathrm{d}e}{\mathrm{d}\theta} \\
&= [1+\alpha(n-1)]D'\left\{(1-\alpha)\theta\frac{\mathrm{d}e_c}{\mathrm{d}\theta} + \alpha\left[(n-m)\frac{\mathrm{d}e_f}{\mathrm{d}\theta} + m\theta\frac{\mathrm{d}e_c}{\mathrm{d}\theta}\right] - \frac{\mathrm{d}e}{\mathrm{d}\theta}\right\} \\
&= \frac{[1+\alpha(m-1)][1+\alpha(n-1)]^2\{(n-m)[1+\alpha(n-1)]D'' - B''(e_f)\}(D')^2(\tilde{\theta}-\theta)}{B''(e_c)B''(e_f) - [1+\alpha(n-1)][(n-m)B''(e_c) + m\theta B''(e_f)]D''},
\end{aligned} \tag{A.72}$$

$$\frac{\mathrm{d}V}{\mathrm{d}\theta} = [1+\alpha(n-1)]\frac{\mathrm{d}W}{\mathrm{d}\theta} = [1+\alpha(n-1)]\left\{(n-m)B'(e_f)\frac{\mathrm{d}e_f}{\mathrm{d}\theta} + mB'(e_c)\frac{\mathrm{d}e_c}{\mathrm{d}\theta} - nD'\frac{\mathrm{d}e}{\mathrm{d}\theta}\right\}$$

$$= [1+\alpha(n-1)]D'\left\{[1+\alpha(n-1)]\left[(n-m)\frac{\mathrm{d}e_f}{\mathrm{d}\theta} + m\theta\frac{\mathrm{d}e_c}{\mathrm{d}\theta}\right] - n\frac{\mathrm{d}e}{\mathrm{d}\theta}\right\}$$

$$= \frac{m[1+\alpha(n-1)]^3\{(n-m)[1+\alpha(n-1)]D'' - B''(e_f)\}(D')^2}{B''(e_c)B''(e_f) - [1+\alpha(n-1)][(n-m)B''(e_c) + m\theta B''(e_f)]D''}$$

$$\cdot\left\{-\frac{(1-\alpha)(n-1)B''(e_f)}{[1+\alpha(n-1)]\{(n-m)[1+\alpha(n-1)]D'' - B''(e_f)\}} + 1 - \theta\right\} \tag{A.73}$$

$$= \frac{m[1+\alpha(n-1)]^3\{(n-m)[1+\alpha(n-1)]D'' - B''(e_f)\}(D')^2}{B''(e_c)B''(e_f) - [1+\alpha(n-1)][(n-m)B''(e_c) + m\theta B''(e_f)]D''}$$

$$\cdot\left\{\frac{(1-\alpha)(n-m)\{[1+\alpha(n-1)]^2D'' - B''(e_f)\}}{[1+\alpha(m-1)][1+\alpha(n-1)]\{(n-m)[1+\alpha(n-1)]D'' - B''(e_f)\}} + \tilde{\theta} - \theta\right\}, \tag{A.74}$$

$$\frac{\mathrm{d}W_f}{\mathrm{d}\theta} = B'(e_f)\frac{\mathrm{d}e_f}{\mathrm{d}\theta} - D'\frac{\mathrm{d}e}{\mathrm{d}\theta} > 0, \tag{A.75}$$

$$\frac{\mathrm{d}W_c}{\mathrm{d}\theta} = B'(e_c)\frac{\mathrm{d}e_c}{\mathrm{d}\theta} - D'\frac{\mathrm{d}e}{\mathrm{d}\theta} = D'\left\{[1+\alpha(n-1)]\theta\frac{\mathrm{d}e_c}{\mathrm{d}\theta} - \frac{\mathrm{d}e}{\mathrm{d}\theta}\right\}$$

$$= -\frac{[1+\alpha(n-1)]^2\{(n-m)[1+\alpha(n-1)]D'' - B''(e_f)\}(D')^2}{B''(e_c)B''(e_f) - [1+\alpha(n-1)][(n-m)B''(e_c) + m\theta B''(e_f)]D''}$$

$$\cdot\left\{\frac{\alpha m(n-m)\{(n-m)[1+\alpha(n-1)]^2D'' - B''(e_f)\}}{[1+\alpha(m-1)][1+\alpha(n-1)]\{(n-m)[1+\alpha(n-1)]D'' - B''(e_f)\}} + \theta - \tilde{\theta}\right\}, \tag{A.76}$$

$$\frac{\mathrm{d}(V_f - V_c)}{\mathrm{d}\theta} = (1-\alpha)\frac{\mathrm{d}(W_f - W_c)}{\mathrm{d}\theta} = (1-\alpha)\left[B'(e_f)\frac{\mathrm{d}e_f}{\mathrm{d}\theta} - B'(e_c)\frac{\mathrm{d}e_c}{\mathrm{d}\theta}\right] > 0. \tag{A.77}$$

(A.70) [(A.73)] yields $\frac{\mathrm{d}V_f}{\mathrm{d}\theta} > 0$ $\left[\frac{\mathrm{d}V}{\mathrm{d}\theta} > 0 \text{ and } \frac{\mathrm{d}W}{\mathrm{d}\theta} > 0\right]$ for $\theta \leq 1$, which implies $V_f < V_i^{\mathrm{BAU}}$ [$V < V^{\mathrm{BAU}}$ and $W < W^{\mathrm{BAU}}$] for $\tilde{\theta} < 1$. (A.71) [(A.74)] yields $\frac{\mathrm{d}V_f}{\mathrm{d}\theta} > 0$ $\left[\frac{\mathrm{d}V}{\mathrm{d}\theta} > 0 \text{ and } \frac{\mathrm{d}W}{\mathrm{d}\theta} > 0\right]$ for $\theta \leq \tilde{\theta}$, which implies $V_f > V_i^{\mathrm{BAU}}$ [$V > V^{\mathrm{BAU}}$ and $W > W^{\mathrm{BAU}}$] for $\tilde{\theta} > 1$. (A.72) yields $\frac{\mathrm{d}V_c}{\mathrm{d}\theta} \gtreqless 0$ for $\theta \gtreqless \tilde{\theta}$, which implies $V_c > V_i^{\mathrm{BAU}}$ for $\tilde{\theta} \neq 1$. (A.75) implies $W_f \gtreqless W_i^{\mathrm{BAU}}$ for $\tilde{\theta} \gtreqless 1$. (A.76) yields $\frac{\mathrm{d}W_c}{\mathrm{d}\theta} < 0$ for $\theta \geq \tilde{\theta}$, which implies $W_c > W_i^{\mathrm{BAU}}$ for $\tilde{\theta} < 1$. Finally, (A.77) implies $V_f \gtreqless V_c$ and $W_f \gtreqless W_c$ for $\tilde{\theta} \gtreqless 1$.

Third suppose $e \leq e^{\mathrm{SO}}$. Then, the right-hand sides of (4) and (13) would be smaller than $nD'(e^{\mathrm{SO}})$, such that the left-hand sides would have to be smaller than $B'(e_i^{\mathrm{SO}})$, implying $e_c, e_f < e_i^{\mathrm{SO}}$ and contradicting $e \leq e^{\mathrm{SO}}$. Thus, $e > e^{\mathrm{SO}}$. □

Proof of Proposition 5

Totally differentiating (4), (13) and $e = me_c + (n-m)e_f$ yields

$$B''(e_f)\mathrm{d}e_f = [1+\alpha(n-1)]D''\mathrm{d}e + (n-1)D'\mathrm{d}\alpha, \tag{A.78}$$

$$B''(e_c)\mathrm{d}e_c = \lambda_e \mathrm{d}e - \lambda_{e_f}\mathrm{d}e_f + \lambda_m \mathrm{d}m + \lambda_\alpha \mathrm{d}\alpha, \tag{A.79}$$

$$\mathrm{d}e = m\mathrm{d}e_c + (n-m)\mathrm{d}e_f + (e_c - e_f)\mathrm{d}m, \tag{A.80}$$

where

$$\begin{aligned}
\lambda_e :=\ & \frac{1+\alpha(n-1)}{1+\alpha(m-1)} mD'' \frac{\alpha(n-m)[1+\alpha(n-1)]D'' - B''(e_f)}{(n-m)[1+\alpha(n-1)]D'' - B''(e_f)} \\
& + \frac{1+\alpha(n-1)}{1+\alpha(m-1)} mD' \frac{(1-\alpha)(n-m)[1+\alpha(n-1)]B''(e_f)}{[(n-m)[1+\alpha(n-1)]D'' - B''(e_f)]^2} D''' > 0 \Longleftarrow D''' \leq 0, \\
\lambda_{e_f} :=\ & \frac{1+\alpha(n-1)}{1+\alpha(m-1)} mD' \frac{(1-\alpha)(n-m)[1+\alpha(n-1)]D''}{[(n-m)[1+\alpha(n-1)]D'' - B''(e_f)]^2} B'''(e_f) \gtreqless 0 \Longleftrightarrow B''' \gtreqless 0, \\
\lambda_m :=\ & \frac{1+\alpha(n-1)}{[1+\alpha(m-1)]^2} D' \frac{1-\alpha}{[(n-m)[1+\alpha(n-1)]D'' - B''(e_f)]^2} \{\alpha(n-m)^2[1+\alpha(n-1)]^2 \\
& \cdot [D'']^2 - [n+\alpha(m^2+n-2m)][1+\alpha(n-1)]D''B''(e_f) + [B''(e_f)]^2]\} > 0, \\
\lambda_\alpha :=\ & \frac{n-m}{[1+\alpha(m-1)]^2} mD' \frac{1}{[(n-m)[1+\alpha(n-1)]D'' - B''(e_f)]^2} \{[1+\alpha(n-1)][1+\alpha(m \\
& -1)][1+2\alpha(n-1)+\alpha^2(n-1)(m-1)][(n-m)[1+\alpha(n-1)]D'' - B''(e_f)]D''(1 \\
& -\tilde{\theta})/(1-\alpha) + [1+\alpha(n-1)][1+\alpha(n-1)(m^2-1)+\alpha(n-1)(m+2)]D''B''(e_f) \\
& + [B''(e_f)]^2\} > 0 \Longleftarrow \tilde{\theta} \leq 1.
\end{aligned}$$

Solving for $\mathrm{d}e_f$, $\mathrm{d}e_c$ and $\mathrm{d}e$ yields

$$\begin{aligned}
&\lambda \mathrm{d}e_f \\
&= [1+\alpha(n-1)]D''[m\lambda_m - (e_f - e_c)B''(e_c)]\mathrm{d}m \\
&\quad - \{(n-1)[m\lambda_e - B''(e_c)]D' - m[1+\alpha(n-1)]D''\lambda_\alpha\}\mathrm{d}\alpha,
\end{aligned} \tag{A.81}$$

$$\begin{aligned}
&\lambda \mathrm{d}e_c \\
&= -\{[1+\alpha(n-1)]D''[(n-m)\lambda_m + (e_c - e_f)\lambda_{e_f}] - B''(e_f)[\lambda_m + (e_c - e_f)\lambda_e]\}\mathrm{d}m \\
&\quad + \{(n-1)[(n-m)\lambda_e - \lambda_{e_f}]D' - \{(n-m)[1+\alpha(n-1)]D'' - B''(e_f)\}\lambda_\alpha\}\mathrm{d}\alpha,
\end{aligned} \tag{A.82}$$

$$\begin{aligned}
&\lambda \mathrm{d}e \\
&= B''(e_f)[m\lambda_m - (e_f - e_c)B''(e_c)]\mathrm{d}m \\
&\quad - \{(n-1)[m\lambda_{e_f} - (n-m)B''(e_c)]D' - mB''(e_f)\lambda_\alpha\}\mathrm{d}\alpha,
\end{aligned} \tag{A.83}$$

where

$$\lambda := [1+\alpha(n-1)]D''[m\lambda_{e_f} - (n-m)B''(e_c)] - [m\lambda_e - B''(e_c)]B''(e_f) > 0$$
$$\Longleftarrow D''' \leq 0, B''' \geq 0.$$

First differentiating (14) with respect to m and using (A.81) and (A.83) yields

$$\begin{aligned}\frac{\mathrm{d}\tilde{\theta}}{\mathrm{d}m} = &-\frac{1}{[1+\alpha(n-1)][1+\alpha(m-1)]\lambda\{(n-m)[1+\alpha(n-1)]D'' - B''(e_f)\}^2 D'} \\ &\cdot \{(n-m)\lambda_m\{(n-m)[1+\alpha(n-1)]D'' - B''(e_f)\}\{[1+\alpha(m-1)]\{(n-m)[1+\alpha \\ &\cdot (n-1)]D'' - B''(e_f)\}^2 B''(e_c) + m^2[1+\alpha(n-1)]\{\alpha(n-m)[1+\alpha(n-1)]D'' \\ &- B''(e_f)\}D''B''(e_f)\} - (e_c - e_f)(1-\alpha)m(n-m)[1+\alpha(n-1)]^2 B''(e_c)\{[B''(e_f)]^2 \\ &\cdot D''' - [1+\alpha(n-1)][D'']^2 B'''(e_f)\}D'\},\end{aligned} \tag{A.84}$$

such that $e_c \geq e_f \Longleftrightarrow \tilde{\theta} \leq 1 \implies \frac{\mathrm{d}\tilde{\theta}}{\mathrm{d}m} > 0$ if $D''' \leq 0, B''' \geq 0$. $\tilde{\theta} \leq 1 \implies \frac{\mathrm{d}\tilde{\theta}}{\mathrm{d}m} > 0$ implies that $\tilde{\theta}(\underline{m}) \geq 1 \implies \tilde{\theta}(\bar{m}) > 1$ for $\bar{m} > \underline{m}$. Thus, (14) implicitly defines $\tilde{m}$ with $m \lesseqgtr \tilde{m} \Longleftrightarrow \tilde{\theta} \lesseqgtr 1$ if $D''' \leq 0, B''' \geq 0$. Using $\tilde{\theta} = 1$ in (14) and solving for m yields (15).

Second differentiating (15) with respect to α yields

$$\begin{aligned}\frac{\mathrm{d}\tilde{m}}{\mathrm{d}\alpha} = &-\frac{(n-1)^2 D''(e^{\mathrm{BAU}})B''(e_i^{\mathrm{BAU}})}{\{[1+\alpha(n-1)]D''(e^{\mathrm{BAU}}) - B''(e_i^{\mathrm{BAU}})\}^2} \\ &- \frac{(n-1)[1+\alpha(n-1)]B''(e_i^{\mathrm{BAU}})D'''(e^{\mathrm{BAU}})}{\{[1+\alpha(n-1)]D''(e^{\mathrm{BAU}}) - B''(e_i^{\mathrm{BAU}})\}^2}\frac{\mathrm{d}e^{\mathrm{BAU}}}{\mathrm{d}\alpha} \\ &+ \frac{(n-1)[1+\alpha(n-1)]D''(e^{\mathrm{BAU}})B'''(e_i^{\mathrm{BAU}})}{\{[1+\alpha(n-1)]D''(e^{\mathrm{BAU}}) - B''(e_i^{\mathrm{BAU}})\}^2}\frac{\mathrm{d}e_i^{\mathrm{BAU}}}{\mathrm{d}\alpha},\end{aligned} \tag{A.85}$$

where $\frac{\mathrm{d}e^{\mathrm{BAU}}}{\mathrm{d}\alpha} = \frac{\mathrm{d}ne_i^{\mathrm{BAU}}}{\mathrm{d}\alpha} < 0$ from Appendix A.1. Thus, $\frac{\mathrm{d}\tilde{m}}{\mathrm{d}\alpha} > 0$ if $D''' \leq 0, B''' \leq 0$. □

Proof of Lemma 2

First (A.81) [(A.83)] yields $\frac{\mathrm{d}e_f}{\mathrm{d}m} > 0$ [$\frac{\mathrm{d}e}{\mathrm{d}m} < 0$] for $e_f \geq e_c \Longleftrightarrow m \geq \tilde{m}$, and (A.82) yields $\frac{\mathrm{d}e_c}{\mathrm{d}m} < 0$ for $e_c \geq e_f \Longleftrightarrow m \leq \tilde{m}$ if $D''' \leq 0, B''' \geq 0$. Furthermore, using (14) in (A.81) yields

$$\begin{aligned}&-(n-1)m\lambda_e|_{D''=0}D' + m[1+\alpha(n-1)]D''\lambda_\alpha \\ &= -\frac{m^2[1+\alpha(n-1)]D'D''}{(1-\alpha)^2(m-1)[1+\alpha(m-1)][(n-m)[1+\alpha(n-1)]D'' - B''(e_f)]^2}\{(1-\alpha)^3(n-1) \\ &\cdot [1+\alpha(n-1)](n-m)^2[D'']^2 + (1-\alpha)[(1-\alpha)^2(n+m-1) + 2\alpha(1-\alpha)nm + \alpha^2 nm] \\ &\cdot (n-m)D''[(n-m)[1+\alpha(n-1)]D'' - B''(e_f)](\tilde{\theta}-1) + [1+\alpha(m-1)][(n-m)[1 \\ &+ \alpha(n-1)]D'' - B''(e_f)]^2(\tilde{\theta}-1)^2\} < 0 \Longleftarrow \tilde{\theta} \geq 1,\end{aligned}$$

such that $\frac{de_f}{d\alpha} < 0$ for $e_f \geq e_c \iff m \geq \tilde{m}$, and (A.83) yields $\frac{de}{d\alpha} < 0$ for $e_c \geq e_f \iff m \leq \tilde{m}$ if $D''' \leq 0, B''' \geq 0$.

Second differentiating V_i and W_i with respect to m and using (4), (6), (13) and (A.81)-(A.83) yields

$$\begin{aligned}\frac{dV_f}{dm} &= [1+\alpha(n-1)]D'\left\{(1-\alpha)\frac{de_f}{dm} + \alpha\left[(n-m)\frac{de_f}{dm} + m\theta\frac{de_c}{dm}\right] - \frac{de}{dm}\right\} \\ &= \frac{[1+\alpha(n-1)]D'}{[1+\alpha(m-1)]\lambda\{(n-m)[1+\alpha(n-1)]D'' - B''(e_f)\}} \\ &\quad \cdot \{(1-\alpha)m\{[1+\alpha(n-1)]^2 D'' - B''(e_f)\}\{(n-m)[1+\alpha(n-1)]D'' - B''(e_f)\}\lambda_m \\ &\quad + (e_f - e_c)\{\alpha m^2\{\alpha(n-m)[1+\alpha(n-1)]D'' - B''(e_f)\}\{[1+\alpha(n-1)]D''\lambda_{e_f} - B''(e_f) \\ &\quad \cdot \lambda_e\}B''(e_f)\} - [1+\alpha(m-1)]\{[1+\alpha(n-m-1)][1+\alpha(n-1)]D'' - B''(e_f)\}\{(n \\ &\quad - m)[1+\alpha(n-1)]D'' - B''(e_f)\}B''(e_c)\},\end{aligned} \tag{A.86}$$

$$\begin{aligned}\frac{dV_c}{dm} &= [1+\alpha(n-1)]D'\left\{(1-\alpha)\theta\frac{de_c}{dm} + \alpha\left[(n-m)\frac{de_f}{dm} + m\theta\frac{de_c}{dm}\right] - \frac{de}{dm}\right\} \\ &= \frac{(e_f - e_c)[1+\alpha(n-1)]D'\{\alpha(n-m)[1+\alpha(n-1)]D'' - B''(e_f)\}}{\lambda\{(n-m)[1+\alpha(n-1)]D'' - B''(e_f)\}} \\ &\quad \cdot \{m\{[1+\alpha(n-1)]D''\lambda_{e_f} - B''(e_f)\lambda_e\} - \{(n-m)[1+\alpha(n-1)]D'' - B''(e_f)\}B''(e_c)\},\end{aligned} \tag{A.87}$$

$$\frac{dW_f}{dm} = D'\left\{[1+\alpha(n-1)]\frac{de_f}{dm} - \frac{de}{dm}\right\}, \tag{A.88}$$

$$\begin{aligned}\frac{dW_c}{dm} &= D'\left\{[1+\alpha(n-1)]\theta\frac{de_c}{dm} - \frac{de}{dm}\right\} \\ &= -\frac{D'}{[1+\alpha(m-1)]\lambda}\{\alpha m(n-m)\{[1+\alpha(n-1)]^2 D'' - B''(e_f)\}\lambda_m + (e_c - e_f)[1+\alpha(m \\ &\quad - 1)]\{[1+\alpha(n-1)]\theta\{[1+\alpha(n-1)]D''\lambda_{e_f} - B''(e_f)\lambda_e\} + B''(e_f)B''(e_c)\}\},\end{aligned} \tag{A.89}$$

such that $\frac{dV_f}{dm} > 0$ for $e_f \geq e_c \iff m \geq \tilde{m}$, and $\frac{dV_c}{dm} \gtreqless 0$ for $e_f \gtreqless e_c \iff m \gtreqless \tilde{m}$ if $D''' \leq 0, B''' \geq 0$. Furthermore, $\frac{de_f}{dm} > 0$ and $\frac{de}{dm} < 0$ implies $\frac{dW_f}{dm} > 0$ for $m \geq \tilde{m}$, and $e_c \geq e_f$ implies $\frac{dW_c}{dm} < 0$ for $m \leq \tilde{m}$.

Third differentiating W with respect to α and using (4), (6), (13), (A.81), (A.82) and (A.83) yields

$$
\begin{aligned}
\frac{\mathrm{d}W}{\mathrm{d}\alpha} &= (n-m)\frac{\mathrm{d}e_f}{\mathrm{d}\alpha} + mB'(e_c)\frac{\mathrm{d}e_c}{\mathrm{d}\alpha} - nD'\frac{\mathrm{d}e}{\mathrm{d}\alpha} \\
&= D'\left\{(n-m)[1+\alpha(n-1)]\frac{\mathrm{d}e_f}{\mathrm{d}\alpha} + m[1+\alpha(n-1)]\theta\frac{\mathrm{d}e_c}{\mathrm{d}\alpha} - n\frac{\mathrm{d}e}{\mathrm{d}\alpha}\right\} \\
&= \frac{(n-1)(D')^2}{\lambda}\left\{\frac{(\theta-1)m^2(n-m)^2(1-\alpha)[1+\alpha(n-1)]^3B''(e_f)D'D'''}{[1+\alpha(m-1)][(n-m)[1+\alpha(n-1)]D''-B''(e_f)]^2}\right. \\
&\quad -(1-\alpha)(n-1)(n-m)B''(e_c) + m\{n-[1+\alpha(n-1)]\theta\}\lambda_{e_f} \\
&\quad + \frac{m^2(n-m)(1-\alpha)[1+\alpha(n-1)]^3(D'')^3}{(n-1)(m-1)^3[1+\alpha(m-1)][(n-m)[1+\alpha(n-1)]D''-B''(e_f)]^2} \\
&\quad \cdot\{(n-1)^2(n-m)^2[1+\alpha(m^2-1)] + (n-1)(n-m)[1+\alpha(m-1)][3n-2m-1+\alpha \\
&\quad \cdot(n-1)(m^2-1)]\psi + \{3(n-m)^2 + 2(m-1)(n-m) + (m-1)^3 + \alpha(n-1)(m-1) \\
&\quad \cdot[(m^2-1)(n-m)\alpha + (m-1)^3\alpha + 4(n-m) + 2(m-1)^2]\}\psi^2 + (n-m)[1+\alpha(m \\
&\quad \left.-1)]\psi^3\}\right\} > 0 \Longleftarrow \tilde{\theta} \geq 1, D''' \leq 0, B''' \geq 0,
\end{aligned}
\tag{A.90}
$$

where $\psi := \frac{(\tilde{\theta}-1)\{(n-m)[1+\alpha(n-1)]D''-B''(e_f)\}}{(1-\alpha)[1+\alpha(n-1)]D''}$, such that $\frac{\mathrm{d}W}{\mathrm{d}\alpha} > 0$ for $e_f \geq e_c \Longleftrightarrow m \geq \tilde{m}$ if $D''' \leq 0, B''' \geq 0$.

Furthermore, differentiating (6) with respect to α and using (A.83) yields

$$
\begin{aligned}
\frac{\mathrm{d}R_f'}{\mathrm{d}\alpha} &= \frac{(n-m)(n-1)D''B''(e_f)}{\{(n-m)[1+\alpha(n-1)]D''-B''(e_f)\}^2} \\
&\quad + \frac{(n-m)[1+\alpha(n-1)]D'''B''(e_f)}{\{(n-m)[1+\alpha(n-1)]D''-B''(e_f)\}^2}\frac{\mathrm{d}e}{\mathrm{d}\alpha} \\
&\quad - \frac{(n-m)[1+\alpha(n-1)]D''B'''(e_f)}{\{(n-m)[1+\alpha(n-1)]D''-B''(e_f)\}^2}\frac{\mathrm{d}e_f}{\mathrm{d}\alpha} \\
&= \frac{(n-m)B''(e_f)}{\lambda\{(n-m)[1+\alpha(n-1)]D''-B''(e_f)\}^2}\{m(n-1)[1+\alpha(n-1)][(D'')^2-D'D'''] \\
&\quad \cdot[\lambda_{e_f} - (n-m)B''(e_c)] - (n-1)D''B''(e_f)[m\lambda_e|_{D'''=0} - B''(e_c)] + m^2(n-m) \\
&\quad \cdot[1+\alpha(n-1)]D'D'''B''(e_f)\{[1+2\alpha(n-1)+\alpha^2(m-1)(n-1)][1+\alpha(n-1)]D'' \\
&\quad -B''(e_f)\}/\{[1+\alpha(m-1)]^2[(n-m)[1+\alpha(n-1)]D''-B''(e_f)]\}\} \\
&\quad - \frac{(n-m)[1+\alpha(n-1)]D''B'''(e_f)}{\{(n-m)[1+\alpha(n-1)]D''-B''(e_f)\}^2}\frac{\mathrm{d}e_f}{\mathrm{d}\alpha},
\end{aligned}
\tag{A.91}
$$

which is negative if $D''' \leq 0$ and $B''' = 0$. Finally, using $D''' = 0$ and $B''' = 0$ yields

$$\frac{\mathrm{d}(1-\alpha)R_f'}{\mathrm{d}\alpha} = \frac{(n-m)D''\{(n-m)[1+\alpha(n-1)]^2D'' + [n-2-2\alpha(n-1)]B''(e_f)\}}{\{(n-m)[1+\alpha(n-1)]D'' - B''(e_f)\}^2}, \tag{A.92}$$

$$(1-\alpha)R_f' - R_f'|_{\alpha=0} = \frac{\alpha(n-m)D''\{(n-m)[1+\alpha(n-1)]D'' + [n-2-\alpha(n-1)]B''(e_f)\}}{\{(n-m)[1+\alpha(n-1)]D'' - B''(e_f)\}\{(n-m)D'' - B''(e_f)\}}, \tag{A.93}$$

such that $(1-\alpha)R_f'$ increases with α for $\alpha \geq 0.5\frac{n-2}{n-1}$, $D''' = 0$ and $B''' = 0$ (sufficient), and $(1-\alpha)R_f'$ is greater than $R_f'|_{\alpha=0}$ for $\alpha \geq \frac{n-2}{n-1}$, $D''' = 0$ and $B''' = 0$ (sufficient). □

Prove of Proposition 6

The equilibrium is defined by

$$a - be_f = [1+\alpha(n-1)]de, \tag{A.94}$$

$$a - be_c = \frac{1+\alpha(n-1)}{1+\alpha(m-1)}mde\left[1-(1-\alpha)\frac{(n-m)[1+\alpha(n-1)]d}{(n-m)[1+\alpha(n-1)]d-b}\right], \tag{A.95}$$

$$e = (n-m)e_f + me_c. \tag{A.96}$$

Solving for e_f, e_c and e yields

$$e_f = \frac{a}{b\omega}\left\{1+\alpha(m-1) + [1+\alpha(n-1)]\{(n-m)[1+\alpha(m-1)] + (1-\alpha)m(m-1)\}\frac{d}{b} - m(n-m)(1-\alpha)[1+\alpha(n-1)]^2\left(\frac{d}{b}\right)^2\right\}, \tag{A.97}$$

$$e_c = \frac{a}{b\omega}\left\{1+\alpha(m-1) - [1+\alpha(n-1)](n-m)\{m-2-2\alpha(m-1)\}\frac{d}{b} + (n-m)^2(1-\alpha)[1+\alpha(n-1)]^2\left(\frac{d}{b}\right)^2\right\} > 0 \Longleftarrow \alpha \geq \frac{1}{2}, \tag{A.98}$$

$$e = \frac{a}{b\omega} n[1+\alpha(m-1)]\left\{1+(n-m)[1+\alpha(n-1)]\frac{d}{b}\right\} > 0, \tag{A.99}$$

where

$$\omega := 1+\alpha(m-1)+[1+\alpha(n-1)]\{2(n-m)[1+\alpha(m-1)]+m^2\}\frac{d}{b}$$
$$+(n-m)[1+\alpha(n-1)]^2\{[1+\alpha(m-1)](n-m)+\alpha m^2\}\left(\frac{d}{b}\right)^2 > 0.$$

Note that $\frac{\partial[}{\partial 2}]\frac{e_c\omega}{1+\alpha(m-1)}m > 0$ for $\beta \le \frac{1}{2}$. For $\alpha = 0$, $\frac{e_c\omega}{1+\alpha(m-1)}$ is minimal at $m = \frac{n+2}{2} + \frac{n-2}{2}\frac{d}{b+d}$, and then $\frac{e_c\omega}{1+\alpha(m-1)}$ is non-negative if and only if $d \le \bar{\bar{d}} := \frac{4b}{n(n-4)}$, which is thus an upper bound for d. It can be shown that $\frac{e_f\omega}{1+\alpha(m-1)}$ is positive for $m \in [2,n]$ and $\alpha \in [0,1]$ if $d \le \bar{\bar{d}}$ and $n \ge 6$. The corresponding Maple file is available on request.

Using (A.97), (A.98) and (A.99) yields

$$V_f = [1+\alpha(n-m-1)]\left[ae_f - \frac{b}{2}e_f^2 - \frac{d}{2}e^2\right] + \alpha m\left[ae_c - \frac{b}{2}e_c^2 - \frac{d}{2}e^2\right]$$
$$= V_c + \frac{a^2}{2b\omega^2}n^2(m-\tilde{m})(1-\alpha)^2[1+\alpha(n-1)]^2\left\{1+[1+\alpha(n-1)]\frac{d}{b}\right\}\left\{m+1\right.$$
$$\left.+\alpha(m-1)+(n-m)[1+\alpha(n-1)][1+\alpha(2m-1)]\frac{d}{b}\right\}\left(\frac{d}{b}\right)^2, \tag{A.100}$$

$$V_c = [1+\alpha(n-m)]\left[ae_f - \frac{b}{2}e_f^2 - \frac{d}{2}e^2\right] + [1+\alpha(m-1)]\left[ae_c - \frac{b}{2}e_c^2 - \frac{d}{2}e^2\right]$$
$$= \frac{a^2}{2b\omega}[1+\alpha(n-1)]\left\{1+\alpha(m-1)-(n-m)\{n+m-2-2\alpha^2(n-1)(m-1)\right.$$
$$\left.+\alpha(nm-3n-3m+4)\}\frac{d}{b}+(n-m)^2(1-\alpha)^2[1+\alpha(n-1)]\left(\frac{d}{b}\right)^2\right\}, \tag{A.101}$$

where $\tilde{m} = \frac{1+n[1+\alpha(n-1)]d/b}{1+[1+\alpha(n-1)]d/b}$.

The internal stability condition reads

$$V_c(m) - V_f(m-1) = \frac{a^2n^2(1-\alpha)^2[1+\alpha(n-1)]^2\left(\frac{d}{b}\right)^2}{2b\omega(m)\omega(m-1)^2}\phi(m), \tag{A.102}$$

where

$$\begin{aligned}\phi(m) :=& -(m-1)[m-3+\alpha(m-2)^2]-(m-1)[1+\alpha(n-1)]\{[(4m^2-12m+4)(n\\ &-m)+m^3-2m^2-4m+2]\alpha^2+[(2m^2-2m-10)(n-m)+2m^3-5m^2-6]\alpha\\ &\cdot(1-\alpha)+[(2(m-3))(n-m)+m^3-3m^2+4m-8](1-\alpha)^2\}\frac{d}{b}-[1+\alpha(n\\ &-1)]^2\{(m-1)[(5m^2-17m+6)(n-m)^2+(3m^3-6m^2-12m+6)(n-m)\\ &+m^3-5m^2+1]\alpha^2+(m-1)[(m^2+2m-17)(n-m)^2+(3m^3-9m^2+4m\\ &-20)(n-m)+2m^3-8m^2+2m-5]\alpha(1-\alpha)+[(m^2-4m+2)(n-m)^2-8\\ &\cdot(m-1)(n-m)-m^2-4m+5](1-\alpha)^2\}\left(\frac{d}{b}\right)^2+(n-m+1)[1+\alpha(n-1)]^3\\ &\cdot\{(m-1)[(2m^2-10m+4)(n-m)^2+(2m^3-8m^2-m+2)(n-m)-2m^2\\ &+m]\alpha^2+(m-1)[(2m-12)(n-m)^2-(2m^2+9)(n-m)-2m^2-1]\alpha(1-\alpha)\\ &-[2(n-m)^2+(m^2-1)(n-m)+m^2-1](1-\alpha)^2\}\left(\frac{d}{b}\right)^3+(n-m)(n-m\\ &+1)^2[1+\alpha(n-1)]^4\{n(2m^2-3m+1)\alpha^2+(3n-2m+1)(m-1)\alpha(1-\alpha)+(n\\ &-m)(1-\alpha)^2\}\left(\frac{d}{b}\right)^4.\end{aligned} \tag{A.103}$$

Substituting $m=\frac{1}{1+\beta}(\tilde{m}+2)+\frac{\beta}{1+\beta}n$, $n=N_7+7$, $\alpha=\frac{1}{1+\gamma}$, $d=\frac{1}{1+\delta}\bar{\bar{d}}$ with $\beta,\gamma,\delta\geq 0$ yields $\phi\left(\frac{1}{1+\beta}(\tilde{m}+2)+\frac{\beta}{1+\beta}n\right)<0$, which implies $\phi(m)<0$ for $m\geq\tilde{m}+2$, $n\geq 7$ and $d\leq\bar{\bar{d}}$. The corresponding Maple file is available on request. Consequently, all coalitions $m\geq\tilde{m}+2$ are internally unstable, while all coalitions $m\leq\tilde{m}$ are externally unstable from Lemma 3. Suppose $\tilde{m}$ is an integer. Then, $m=\tilde{m}$ is externally unstable and $m=\tilde{m}+2$ is internally unstable, $m=\tilde{m}+1$ is internally and externally stable from Lemma 3. Suppose $\tilde{m}$ is not an integer. Then, $m=\lfloor\tilde{m}\rfloor$ is externally unstable and $m=\lfloor\tilde{m}+3\rfloor$ is internally unstable, such that $m=\lfloor\tilde{m}+1\rfloor$ is internally stable from Lemma 3 and $m=\lfloor\tilde{m}+2\rfloor$ is externally stable. If $m=\lfloor\tilde{m}+1\rfloor$ is externally stable [unstable], then $m=\lfloor\tilde{m}+2\rfloor$ is internally stable [unstable], such that some unique coalition $m\in(\tilde{m},\tilde{m}+2)$ is stable. This proves the first bullet of the proposition. Furthermore, $\frac{\partial\tilde{m}}{\partial d}=\frac{(n-1)[1+\alpha(n-1)]}{b\{1+[1+\alpha(n-1)]d/b\}^2}>0$, such that substituting $d=\bar{\bar{d}}$ into $\tilde{m}$ yields an upper bound $\bar{\tilde{m}}$:

$$\tilde{m}\leq\bar{\tilde{m}}:=\frac{n[n+4\alpha(n-1)]}{(n-2)^2+4\alpha(n-1)}, \tag{A.104}$$

where

$$\frac{\partial\bar{\tilde{m}}}{\partial\alpha}=\frac{4n(n-1)^2(n-4)}{[(n-2)^2+4\alpha(n-1)]^2}\gtreqless 0\iff n\gtreqless 4, \tag{A.105}$$

$$\frac{\partial \bar{\tilde{m}}}{\partial n} = \frac{16\alpha(n-1)^2 - 8\alpha(n-1)(n-2) - 4n(n-2)}{[(n-2)^2 + 4\alpha(n-1)]^2} \gtreqless 0$$
$$\iff \alpha \gtreqless \frac{n-2+\sqrt{(n-2)(5n-2)}}{4(n-1)} \in [0.576, 0.809]. \tag{A.106}$$

Thus, $\bar{\tilde{m}}$ is minimal for $\alpha = 0$ and $n = 7$ with $\bar{\tilde{m}} = 1.96$, and it is maximal for $\alpha = 1$ and $n \to \infty$ with $\bar{\tilde{m}} = 5$. $m \in (\tilde{m}, \tilde{m}+2)$ and $\tilde{m} \leq \bar{\tilde{m}}$ then imply $m \in \{2,3\}$ for $\alpha = 0$ and $m \in \{2,3,4,5,6\}$ for $\alpha > 0$. This proves the second bullet of the proposition and the fourth bullet of the proposition, respectively. Furthermore,

$$\begin{aligned}
&\phi(3)|_{\alpha=0} \\
&= -8\frac{d}{b} + (n^2+10n-23)\left(\frac{d}{b}\right)^2 + 2(n-1)^2(n-2)\left(\frac{d}{b}\right)^3 + (n-2)^2(n-3)^2\left(\frac{d}{b}\right)^4 \\
&= -\frac{d}{16b^4}\{(N_{26}^6 + 122N_{26}^5 + 5951N_{26}^4 + 145224N_{26}^3 + 1778248N_{26}^2 + 8867520N_{26} \\
&\quad + 2064240)d^3 + 2n(n-4)(2N_{26}^4 + 174N_{26}^3 + 5587N_{26}^2 + 78148N_{26} + 399316)(\bar{\tilde{d}} - d)d^2 \\
&\quad + n^2(n-4)^2(5N_{26}^2 + 226N_{26} + 2519)(\bar{\tilde{d}} - d)^2 d + 2n^3(n-4)^3(\bar{\tilde{d}} - d)^3\},
\end{aligned} \tag{A.107}$$

such that $m = 3$ is internally unstable for $\alpha = 0$, $n \geq 26$ and $d \leq \bar{\tilde{d}}$. Finally,

$$\begin{aligned}
\phi(3) = &-2\alpha - 2[1+\alpha(n-1)][4+(2n-11)\alpha + (2n-6)\alpha^2]\frac{d}{b} + [1+\alpha(n-1)]^2[n^2+10n \\
&- 23 + (2n^2-28n+68)\alpha - (3n^2-24n+29)\alpha^2]\left(\frac{d}{b}\right)^2 + (n-2)[1+\alpha(n-1)]^3 \\
&\cdot [2(n-1)^2 + (8n^2-10n-20)\alpha + (n-3)(6n-26)\alpha^2]\left(\frac{d}{b}\right)^3 + (n-3)(n-2)^2 \\
&\cdot [1+\alpha(n-1)]^4[n-3+4(n-1)\alpha + (5n+7)\alpha^2]\left(\frac{d}{b}\right)^4,
\end{aligned} \tag{A.108}$$

such that $\phi(3)$ decreases with d/b, and increases with $(d/b)^3$ and $(d/b)^4$, which implies that $\phi(3)$ is positive if and only if d/b is greater than some unique threshold $\underline{d/b} = [\arg \phi(3) = 0]$. Figure 6 shows that the derivative of this threshold with respect to α is negative for $n \in [7, 25]$. This proves the third bullet of the proposition. □

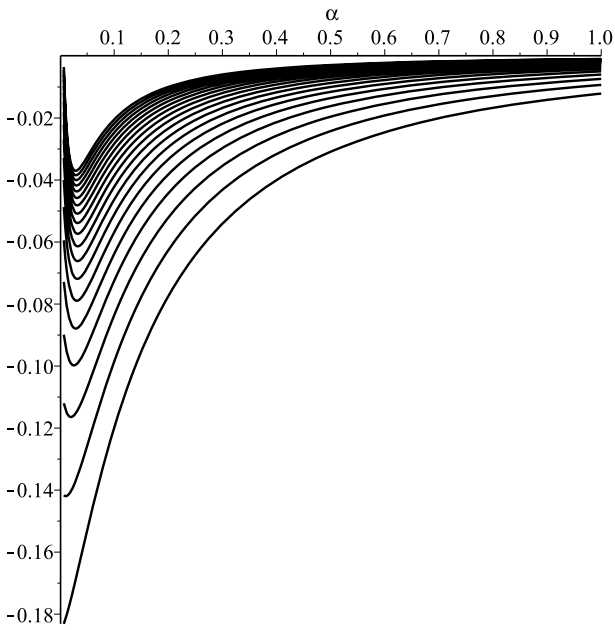

Fig. 6 Derivative of $\overline{d/b}$ with respect to α for $n \in [7, 25]$

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10640-024-00885-8.

Acknowledgements I would like to thank Marc Lenders, Julia Kommritz and participants of the EAERE 2023 and the VfS 2023 for helpful comments. Moreover, the comments from three anonymous reviewers significantly improved this article.

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Conflict of interests The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. We have not received any financial support for the conduct of the research and/or preparation of the article from any third party.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Alger I, Weibull JW (2010) Kinship, incentives, and evolution. Am Econ Rev 100(4):1725–1758

Alger I, Weibull JW (2013) Homo Moralis-preference evolution under incomplete information and assortative matching. Econometrica 81(6):2269–2302

Altamirano-Cabrera J-C, Finus M (2006) Permit trading and stability of international climate agreements. J Appl Econ 9(1):19–47
Andre P, Boneva T, Chopra F, Falk A (2024) Misperceived social norms and willingness to act against climate change. SAFE Working Paper No. 414
Andreoni J (1990) Impure altruism and donations to public goods: a theory of warm-glow giving. Econ J 100(401):464–477
Andreoni J, Harbaugh WT, Vesterlund L (2010) Altruism in experiments. In: Durlauf SN, Blume LE (eds) Behavioural and experimental economics. The New Palgrave Economics Collection, Palgrave Macmillan, London, pp 6–13
Bakalova I, Eyckmans J (2019) Simulating the impact of heterogeneity on stability and effectiveness of international environmental agreements. Eur J Oper Res 277(3):1151–1162
Balliet D, Junhui W, De Dreu CKW (2014) Ingroup favoritism in cooperation: a meta-analysis. Psychol Bull 140(6):1556–1581
Barrett S (1994) Self-enforcing international environmental agreements. Oxf Econ Pap 46:878–894
Barrett S (1997) Heterogeneous international environmental agreements. In: Carraro C (ed) International environmental negotiations: strategic policy issues. Edward Elgar, Cheltenham, pp 9–25
Barrett S (2013) Climate treaties and approaching catastrophes. J Environ Econ Manag 66(2):235–250
Bayramoglu B, Finus M, Jacques J-F (2018) Climate agreements in a mitigation-adaptation game. J Public Econ 165:101–113
Becker GS (1974) A theory of social interactions. J Polit Econ 82(6):1063–1093
Becker GS (1981) Altruism in the family and selfishness in the market place. Economica 48(189):1–15
Bolton GE, Ockenfels A (2000) ERC: a theory of equity, reciprocity, and competition. Am Econ Rev 91(1):166–193
Botteon M, Carraro C (2001) Environmental coalitions with heterogeneous countries: Burden-sharing and carbon leakage. In: Ulph A (ed) Environmental Policy, International Agreements, and International Trade. Oxford University Press, New York
Buchholz W, Haupt A, Peters W (2005) International environmental agreements and strategic voting. Scand J Econ 107(1):175–195
Buchholz W, Peters W, Ufert A (2018) International environmental agreements on climate protection: a binary choice model with heterogeneous agents. J Econ Behav Organ. 154:191–205
Carbone JC, Helm C, Rutherford TF (2009) The case for international emission trade in the absence of cooperative climate policy. J Environ Econ Manag 58(3):266–280
Carraro C, Siniscalco D (1993) Strategies for the international protection of the environment. J Public Econ 52(3):309–328
Cheikbossian G (2021a) Evolutionarily stable in-group altruism in intergroup conflict over (local) public goods. Games Econ Behavr 127:206–226
Cheikbossian G (2021b) The evolutionary stability of in-group altruism in productive and destructive group contests. J Econ Behav Organ 188:236–252
D'Aspremont C, Jacquemin A, Gabszewicz JJ, Weymark JA (1983) On the stability of collusive price leadership. Can J Econ 16(1):17–25
Daube M (2019) Altruism and global environmental taxes. Environ Resource Econ 73(4):1049–1072
Diamantoudi E, Sartzetakis ES (2006) Stable international environmental agreements: an analytical approach. J Public Econ Theory 8(2):247–263
Dietz T, Fitzgerald A, Shwom R (2005) Environmental values. Annu Rev Environ Resour 30:335–372
Eckert H, Nkuiya B (2022) Damage sensitivity and stability in international environmental agreements. Oxf Econ Pap 74(4):1063–1076
Eichner T, Pethig R (2022) International environmental agreements when countries behave morally. CESifo Working Paper No. 10090
Engler D, Gutsche G, Simixhiu A, Ziegler A (2022) Social norms and individual climate protection activities: a framed field experiment for Germany. MAGKS Joint Discussion Paper Series in Economics No. 30-2022
Fehr E, Schmidt KM (1999) A theory of fairness, competition, and cooperation. Q J Econ 114(3):817–868
Finus M (2001) Game theory and international environmental cooperation. Edward Elgar, Cheltenham and Northampton
Finus M, Furini F, Rohrer AV (2021) The efficacy of international environmental agreements when adaptation matters: Nash-Cournot vs. Stackelberg leadership. J Environ Econ Manag 109:102461
Finus M, Furini F, Rohrer AV (2021b) International environmental agreements and the paradox of cooperation: revisiting and generalizing some previous results. Graz Economics Papers No. 2021-05

Finus M, Furini F, Rohrer AV (2023) The Stackelberg vs. Nash-Cournot folk-theorem in international environmental agreements. Econ Lett 234:111481
Finus M, McGinty M (2019) The anti-paradox of cooperation: diversity may pay! J Econ Behav Organ 157:541–559
Fuentes-Albero C, Rubio SJ (2010) Can international environmental cooperation be bought? Eur J Oper Res 202(1):255–264
Goussebaïle A, Bommier A, Goerger A, Nicolaï J-P (2023) Altruistic foreign aid and climate change mitigation. Environ Resource Econ 84(1):219–239
Graziosi GR (2009) On the strategic use of representative democracy in international agreements. J Public Econ Theory 11(2):281–296
Grüning C, Peters W (2010) Can justice and fairness enlarge international environmental agreements? Games 1(2):137–158
Helm C (2003) International emissions trading with endogenous allowance choices. J Public Econ 87(12):2737–2747
Hjerpe M, Löfgren Åsa, Linnér B, Hennlock M, Sterner T, Jagers SC (2011) Common ground for effort sharing? Preferred principles for distributing climate mitigation efforts. University of Gothenburg Working Papers in Economics No. 491
Hoel M (1996) Should a carbon tax be differentiated across sectors? J Public Econ 59(1):17–32
Holtsmark B, Weitzman ML (2020) On the effects of linking cap-and-trade systems for CO_2 emissions. Environ Resource Econ 75(3):615–630
Holtsmark K, Midttømme K (2021) The dynamics of linking permit markets. J Public Econ 198:104406
IPCC (2021) Climate change 2021: the physical science basis. Online at: https://www.ipcc.ch/report/ar6/wg1
Karp L, Simon L (2013) Participation games and international environmental agreements: a non-parametric model. J Environ Econ Manag 65(2):326–344
Kesternich M, Löschel A, Ziegler A (2021) Negotiating weights for burden sharing rules in international climate negotiations: an empirical analysis. Environ Econ Policy Stud 23:309–331
Kotchen MJ, Moore MR (2007) Private provision of environmental public goods: household participation in green-electricity programs. J Environ Econ Manag 53(1):1–16
Lades LK, Laffan K, Weber TO (2021) Do economic preferences predict pro-environmental behaviour? Ecol Econ 183:106977
Lange A (2006) The impact of equity-preferences on the stability of international environmental agreements. Environ Resour Econ 34(2)
Lange A, Vogt C (2003) Cooperation in international environmental negotiations due to a preference for equity. J Public Econ 87(9–10):2049–2067
Lange A, Vogt C, Ziegler A (2007) On the importance of equity in international climate policy: an empirical analysis. Energy Econ 29(3):545–562
Lange A, Schwirplies C (2017) (un)fair delegation: exploring the strategic use of equity rules in international climate negotiations. Environ Resour Econ 67:505–533
Lessmann K, Marschinski R, Finus M, Kornek U, Edenhofer O (2014) Emissions trading with non-signatories in a climate agreement-an analysis of coalition stability. Manch Sch 82:82–109
Loeper A (2017) Cross-border externalities and cooperation among representative democracies. Eur Econ Rev 91:180–208
Markusen JR (1975) International externalities and optimal tax structures. J Int Econ 5(1):15–29
McGinty M (2007) International environmental agreements among asymmetric nations. Oxf Econ Pap 59(1):45–62
McGinty M (2020) Leadership and free-riding: decomposing and explaining the paradox of cooperation in international environmental agreements. Environ Resour Econ 77:449–474
Meulemann M, Ziegler A (2015) The role of burden sharing rules in international climate negotiations. In: Max M (ed) Fairness and components of architectures in international climate negotiations. ETH Zürich, Zürich, pp 19–48
Nkuiya B (2020) Stability of international environmental agreements under isoelastic utility. Resour Energy Econ 59:101128
Nkuiya B, Marrouch W, Bahel E (2015) International environmental agreements under endogenous uncertainty. J Public Econ Theory 17(5):752–772
Nyborg K (2000) Homo economicus and homo politicus: interpretation and aggregation of environmental values. J Econom Behav Organ 42(3):305–322
Nyborg K (2018) Reciprocal climate negotiators. J Environ Econ Manag 92:707–725
Pavlova Y, De Zeeuw A (2013) Asymmetries in international environmental agreements. Environ Dev Econ 18(1):51–68

Pollak RA (1988) Tied transfers and paternalistic preferences. Am Econ Rev 78(2):240–244
Rabin M (1993) Incorporating fairness into game theory and economics. Am Econ Rev 83(5):1281–1302
Ricke K, Drouet L, Caldeira K, Tavoni M (2018) Country-level social cost of carbon. Nat Clim Chang 8(10):895–900
Roemer JE (2015) Kantian optimization: a microfoundation for cooperation. J Public Econ 127:45–57
Rogna M, Vogt CJ (2022) Optimal climate policies under fairness preferences. Clim Change 174(25):1–20
Schwalm CR, Glendon S, Duffy PB (2020) RCP8.5 tracks cumulative CO_2 emissions. Proc Natl Acad Sci 117(33):19656–19657
Segendorff B (1998) Delegation and threat in bargaining. Games Econom Behav 23(2):266–283
Spycher S, Winkler R (2022) Strategic delegation in the formation of modest international environmental agreements. Eur Econ Rev 141:103963
Steg L (2016) Values, norms, and intrinsic motivation to act proenvironmentally. Annu Rev Environ Resour 41:277–292
Tajfel H, Turner JC, Austin WG, Worchel S (1979) An integrative theory of intergroup conflict. In: Austin WG, Worchel S (eds) The social psychology of intergroup relations. Brooks/Cole, Monterey, pp 33–37
The World Bank (2023) Carbon pricing dashboard. Online at: https://carbonpricingdashboard.worldbank.org
Ulph A, Ulph D (2023) International cooperation and Kantian moral behaviour—Complements or substitutes? University of Manchester Economics Working Paper No. EDP-2302
UN (2015) Paris agreement. Online at: https://unfccc.int/process-and-meetings/the-paris-agreement/the-paris-agreement
UN (2023) 'Emissions gap report', Online at: https://www.unep.org/resources/emissions-gap-report-2023
van der Pol T, Weikard H-P, van Ierland E (2012) Can altruism stabilise international climate agreements? Ecol Econ 81:112–120
Vogt C (2016) Climate coalition formation when players are heterogeneous and inequality averse. Environ Resource Econ 65(1):33–59
Yamagishi T, Kiyonari T (1999) Bounded generalized reciprocity: in-group boasting and in- group favoritism. Adv Group Process 16:161–197
Yu S, Yinhao W (2022) Implications of carbon trade with endogenous permits for post-Paris climate agreements. Clim Change Econ 13(4):2250006
Ziegler A (2020) Heterogeneous preferences and the individual change to alternative electricity contracts. Energy Econ 91:104889

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.