European Economic Review 141 (2022) 103963

Contents lists available at ScienceDirect

European Economic Review

journal homepage: www.elsevier.com/locate/eer

Strategic delegation in the formation of modest international environmental agreements

Sarah Spycher *, Ralph Winkler

Department of Economics and Oeschger Centre for Climate Change Research, University of Bern, Schanzeneckstrasse 1, CH-3012 Bern, Switzerland

ARTICLE INFO

JEL classification:
Q54
Q58
C72
D62
H41
P16

Keywords:
International climate policy
Coalition formation game
Political economy
Strategic delegation
Strategic voting

ABSTRACT

We reassess the well-known "narrow-but-deep" versus "broad-but-shallow" trade-off in international environmental agreements (IEAs), taking into account the principal–agent relationship induced by the hierarchical structure of international policy. To this end, we expand the modest coalition formation game, in which countries first decide on whether to join an agreement and then decide on emissions by a strategic delegation stage. In the weak delegation game, principals first decide whether to join an IEA, then delegate the domestic emission choices to an agent. Finally, agents in all countries decide on emissions. In countries not joining the IEA, agents choose emissions to maximize their own payoff, while agents of countries joining the IEA set emissions to internalize some exogenously given fraction γ of the externalities that own emissions cause on all members of the IEA. In the strong delegation game, principals first delegate to agents, who then decide on membership and emissions. We find that strategic delegation crowds out all efforts to increase coalition sizes by less ambitious agreements in the weak delegation game, while in the strong delegation game the first-best from the principals' point of view can be achieved.

1. Introduction

Despite the COVID-19 pandemic, the mitigation of anthropogenic climate change remains one of the most important challenges humanity currently faces. On the positive side, there is a widespread consensus on the long-term policy goal that the increase of the average surface temperature should be contained well below 2 °C compared to the pre-industrial level. This has been formalized in the Paris Agreement in December 2015, which was widely acclaimed by many observers and politicians as a diplomatic breakthrough in international climate policy. On the negative side, we observe little progress in climate change mitigation: In almost all countries, current greenhouse gas emissions are above the agreed upon pledges and even complying with these pledges would not achieve the 2 °C target.

Thus, the Paris Agreement seems to suffer from the well-known "narrow-but-deep" versus "broad-but-shallow" trade-off that is omnipresent in the provision of global public goods, due to the absence of a supranational authority that can enforce cooperation.[1] According to this trade-off, agreements are either ambitious in their effective public good provision but only consist of a small number of participating parties ("narrow-but-deep"), or supported by many or even all involved parties, which comes at a sharp reduction in the effective provision of the public good of each signatory ("broad-but-shallow"). While the "broad-but-shallow"-agreements often beat the "narrow-but-deep"-agreements in the effective aggregate provision of the global public good, as the

* Corresponding author.
E-mail addresses: sarah.spycher@vwi.unibe.ch (S. Spycher), mail@ralph-winkler.de (R. Winkler).

[1] See, for example, Schmalensee (1998), Barrett (2002), Aldy et al. (2003), Finus and Maus (2008) and Harstad (2020).

https://doi.org/10.1016/j.euroecorev.2021.103963
Received 1 December 2020; Received in revised form 28 October 2021; Accepted 29 October 2021
Available online 19 November 2021

reduction in the provision of each signatory is overcompensated by more signatories, they usually fall considerably short of the globally efficient outcome.

In this paper, we reassess the "narrow-but-deep" versus "broad-but-shallow" trade-off in the context of strategic delegation. That is, we depart from the usual assumption that individual countries are represented by a single benevolent decision maker, for example a government, acting in the best interest of the country as a whole. Instead, we account for the "hierarchical structure" of international (environmental) policy. By hierarchical we mean that political decisions in modern societies are not made by a single – let alone benevolent – decision maker. For example, representative democracies typically feature a chain of delegation from voters to those who govern (Strøm, 2000): (i) from voters to elected representatives, (ii) from legislators to the executive branch (head of government), (iii) from the head of government to the heads of different executive departments, and (iv) from these heads to civil servants. In all these situations, one party (an agent) acts on behalf of another (the principal), because the principal either lacks the information or skills of the agent, or simply the time. An additional reason for delegation is that the choice of an agent with certain preferences enables the principal to credibly commit to a particular policy (e.g., Perino, 2010). In this case, the principal delegates *strategically*, i.e., chooses an agent who exhibits preferences that differ from her own.

In our analysis, we start with the analytical framework presented in Finus and Maus (2008), which we amend in two dimensions: First, Finus and Maus (2008) introduce a "modesty" parameter γ into a standard coalition formation game, which represents the fraction of externalities within the coalition that coalition members internalized.[2] While being a parsimonious and analytically tractable deviation from the standard coalition formation game that successfully produces the "narrow-but-deep" versus "broad-but-shallow" trade-off, it raises serious concerns with respect to compliance, as it is not in the best interest of member countries to behave as postulated. To circumvent any issues of compliance, we present an institutional microfoundation based on the international permit market with refunding mechanism introduced by Gersbach and Winkler (2011). In this set-up all agents make decisions such as to maximize their own welfare, i.e., the mechanism is self-enforcing, yet, the outcome in the subgame perfect equilibrium exactly matches the outcome as postulated by Finus and Maus (2008).

Second, we add an additional strategic delegation stage to the model set-up in Finus and Maus (2008). In doing so, we distinguish two institutional settings, depending on how much decision power the principal surrenders to the agent: In the *weak delegation* game the principals in all countries decide in the first stage on whether to join an international environmental agreement (IEA). In the second stage, the principals in all countries select agents, who are in charge of the domestic emission level choices in the third stage. In the *strong delegation* game, the first two stages are interchanged: In the first stage the principals in all countries select agents, who then decide on membership status of the IEA in the second stage and also on emission levels in the third stage.

We find several important and new results. First, with respect to strategic delegation, we show that there are two different motives to strategically delegate: Principals in all countries have an incentive to delegate to agents who exhibit a lower evaluation of environmental damages than themselves. This motive stems from the strategic substitutability of emission choices and is well understood in the literature. By delegating to a "browner" agent, the principal can commit her country to high emission levels, to which the best response of all other countries is to reduce their own emission levels. The second motive is only present for principals of member countries. They have an incentive to delegate to agents that have higher evaluations of the environmental damages than they have themselves. By delegating to a "greener" agent, the principals can increase the "effective" environmental damage (i.e., the environmental damage as measured from their own perspective) of their own country that is internalized by the other member countries. This motive is only present if member countries do not fully internalize the externalities within the coalition, i.e., if the agreement is "shallow". In fact, principals try to counteract any deviation from an ambitious agreement. In the weak delegation framework, where principals know their membership status at the time of delegation, they succeed to fully crowd out any attempt of a modest agreement by choosing an equally "greener" agent, i.e., if, for example, the degree of modesty drops from 1 to 0.5, principals react by delegating to agents who are twice as concerned about the environmental damage. Thus, from the principals' point of view the resulting outcome is as if the agreement was "deep". In the strong delegation game, where principals at the time of delegation do not know whether they end up as coalition members, they can only partially compensate lower degrees of ambition by delegating to "greener" agents. To the best of our knowledge, this second delegation motive is unique to the modest coalition formation model with strategic delegation.

Second, our main result addresses the "narrow-but-deep" versus "broad-but-shallow" trade-off. In fact, we show that this trade-off does not exist in our strategic delegation coalition formation model. In the weak delegation game, any deviation from "narrow-but-deep" is perfectly crowded out by the delegation choices of principals in member countries. Thus, there is no alternative to "narrow-but-deep" agreements. In the strong delegation game, we find, similar to the existing literature on this trade-off, there always exists a sufficiently small γ such that all countries become members of the agreement. Correctly anticipating that the grand coalition forms, the principals know in this case with certainty that they will become member countries and can, as in the weak delegation set-up, fully compensate any attempt of a shallow agreement by delegating to a correspondingly "greener" agent. However, as the membership decision is made by the agents, the grand coalition forms, due to an appropriately small γ, i.e., from the agents' point of view the agreement is "shallow" enough such that full participation is in the best interest of all agents in the second stage. Yet, from the principals' perspective, the agreement actually implements the first-best. Thus, the strong delegation game allows for "broad-and-deep" agreements, in which all countries participate in the agreement and, from the principals' point of view, the globally efficient level of the public good is provided.

[2] Note that the modesty parameter, which in our notation is denoted by 'γ', was called 'α' in Finus and Maus (2008).

We believe that particularly our main result has important consequences for the future design of IEAs. Our analysis shows that strategic delegation is not necessarily an impediment to successful international environmental cooperation. In the right institutional setting, strategic delegation can act as the necessary commitment device for the principals to overcome the free-riding incentives of global public good provision.

Our paper combines two different strands of literature. The first strand is the literature on strategic delegation which emerged in the Industrial Organization (IO) literature analyzing the delegation of managerial decisions from shareholders to chief executive officers (for an excellent survey see Kopel and Pezzino, 2018). Subsequently, the concept of strategic delegation found its way into the literature on negotiation and cooperation (Crawford and Varian, 1979; Sobel, 1981; Jones, 1989; Burtraw, 1992, 1993; Segendorff, 1998), where it has been utilized in various contexts with inter-agent spillovers, such as environmental policy or the provision of public goods more generally.[3]

Siqueira (2003), Buchholz et al. (2005), Roelfsema (2007) and Hattori (2010) analyze strategic voting in the context of environmental policy. Siqueira (2003) and Buchholz et al. (2005) both find that voters' selection of agents is biased toward politicians who are less "green" than the median voter. By electing a "browner" politician, the home country commits itself to a lower tax on pollution, shifting the burden of a cleaner environment to the foreign country. By contrast, Roelfsema (2007) accounts for emissions leakage through shifts in production and finds that median voters may delegate to politicians who place greater weight on environmental damage than they do themselves, whenever their preferences for the environment relative to their valuation of firms' profits are sufficiently strong. However, this result breaks down in the case of perfect pollution spillovers, such as the emission and diffusion of greenhouse gases as in our paper. Hattori (2010) allows for different degrees of product differentiation and alternative modes of competition, i.e., competition on quantities but also on prices. His general finding is that, when the policy choices are strategic substitutes (complements), a less (more) "green" policy maker is elected in the non-cooperative equilibrium. As in Siqueira (2003) and Roelfsema (2007), the agents selected by the principals in our model do not engage in bargaining but rather set environmental policies non-cooperatively.

Strategic delegation in the provision of public goods with cross-border externalities more generally has been examined by Kempf and Rossignol (2013) and Loeper (2017). The authors of the former paper show that any international agreement that is negotiated by national delegates involves higher public good provisions than in the case of non-cooperative policies, taking feasibility, efficiency and equity constraints into account. In their model, the choice of delegates is highly dependent on the distributive characteristics of the proposed agreement. Loeper (2017) proves that whether cooperation between national delegates is beneficial only depends on the type of public good considered and, more specifically, on the curvature of the demand for the public good but not on voters' preferences, the magnitude of the cross-border externalities, nor the size, bargaining power or efficiency of each country in providing the public good. Another strand of this literature deals with the provision of public goods in federations that are characterized by fiscal arrangements or different majoritarian rules; see, e.g., Besley and Coate (2003), Redoano and Scharf (2004), Dur and Roelfsema (2005), Harstad (2010) and Christiansen (2013).

The second strand is the literature on two-stage coalition formation games. For an overview, see the excellent surveys by Barrett (2003), Finus (2001), Wagner (2001) and de Zeeuw (2015). In general, these models draw a pessimistic picture for successful international cooperation: whenever the gains from cooperation would be large, stable coalition sizes are small and, thus, coalitions achieve little (e.g., Carraro and Siniscalco, 1993 and Hoel, 1992).[4] The main idea of the two-stage coalition formation game by Finus and Maus (2008), which we take as basis for our model, is that the coalition does not necessarily internalize all externalities from emissions within the coalition but may opt for a more modest goal, i.e., to internalize only some fraction γ of the environmental damages within the coalition. They show that more modest agreements have higher membership sizes. Although each member of the coalition in a modest agreement emits more than members in an ambitious agreement with equal membership size, this increase in emissions is often outweighed by the larger number of members, who – even in a modest agreement – emit less than non-members of the coalition. Harstad (2020) finds a similar result in a dynamic model that can account for a variety of different empirical observations of international environmental agreements. In contrast to Finus and Maus (2008), he provides a microfoundation for the "narrow-but-deep" versus "broad-but-shallow" trade-off that is inspired by the pledge-and-review mechanism of the Paris Agreement. The decisive difference between these two papers and our paper is that we, in addition, account for the hierarchical structure of international environmental policy by introducing a strategic delegation stage.

From a political economy perspective, the papers most closely related to ours are Marchiori et al. (2017), Hagen et al. (2020), Köke and Lange (2017) and Battaglini and Harstad (2020). All these papers (and ours, too) have in common that they analyze the formation of an IEA in an political economy model, i.e., they explicitly discuss the influence of the interplay of domestic and international climate policy on the prospects of international environmental cooperation. Marchiori et al. (2017) and Hagen et al. (2020) investigate the influence of legislative lobbying, as modeled by a common agency framework, on the formation of IEAs. In a strategic voting model with uncertain median voter preferences Köke and Lange (2017) analyze the impact of ratification constraints on the outcome of IEAs. Battaglini and Harstad (2020) show that the political competition for reelection of an incumbent government with a rival party may have an important impact on the design and the effectiveness of IEAs. In contrast to these papers, we consider a coalition formation game in a strategic delegation framework similar to Habla and Winkler (2018).

[3] It is also worth mentioning that strategic delegation is labeled as *strategic voting* whenever the principal is the electorate or, more precisely, the median voter and the elected government is the agent (Persson and Tabellini, 1992).

[4] However, Karp and Simon (2013) show that this may not necessarily be true.

2. The model

We consider a set $I = \{1, \dots, n\}$ of $n \geq 2$ a priori identical countries. In each country $i \in I$, emissions e_i imply country-specific benefits from productive activities, characterized by a concave quadratic benefit function $B(e_i)$, while global emissions $E = \sum_{i \in I} e_i$ cause convex quadratic damages, $D(E)$. Whenever possible, we formulate our results in terms of generic benefit and damage functions, taking the assumed properties into account. When specific benefit and damage functions are necessary to derive unambiguous conclusions, we employ the following:

$$B(e_i) = \beta e_i \left(\epsilon - \frac{1}{2} e_i\right) , \quad D(E) = \frac{\delta}{2} E^2 , \tag{1}$$

where ϵ denotes the business-as-usual emissions of a country that accrued if no emission reductions because of environmental damages were beneficial. The parameter β measures emissions efficiency, i.e., how much GDP a country can produce per unit of emissions, while the parameter δ is a measure of the environmental damage (in monetary terms) that is caused by global emissions.

2.1. Agency structure

In each country $i \in I$, there is a principal, whose utility is given by:

$$U_i = B(e_i) - \theta_{i,P} D(E) . \tag{2}$$

Without loss of generality, we normalize the principal's preference parameter to unity, i.e., $\theta_{i,P} \equiv 1$. In addition, there is a continuum of agents of mass one in each country i, whose utilities are given by:

$$V_i = B(e_i) - \theta_i D(E) , \tag{3}$$

where θ_i is a preference parameter that is continuously distributed on the bounded interval $[0, \theta^{\max}]$. We assume that the boundary $\theta^{\max}$ is such that (i) the principals' preferences are represented in the continuum of agents, i.e., $\theta^{\max} \geq 1$, and (ii) the principal can always find her preferred agent within the continuum of agents.

Our preference specification implies that in each country, all agents and the principal have equal stakes in the benefits from productive activities, but differ with respect to environmental damages. This may be either because damages are heterogeneously distributed or because the monetary valuation of homogeneous physical environmental damages differs. We assume that all individuals (principals and agents) are selfish in the sense that they maximize their respective utilities, i.e., the principal in country i chooses *her* actions to maximize U_i, while each agent in country i makes decisions to maximize *his* utility V_i. In addition, we assume that preference parameters of all individuals are common knowledge. Thus, we abstract from all issues related to asymmetric information.

2.2. Modest international environmental agreements

We model the hierarchical structure of climate policy as a coalition formation game similar to the model presented by Finus and Maus (2008), which we amend by a strategic delegation stage. In the standard coalition formation game, all countries simultaneously and non-cooperatively decide in the first stage whether to join an agreement. Throughout the paper, we shall call countries that join the agreement "members" and the remaining countries "non-members" or "free-riders". In the second stage, all countries simultaneously set emission levels. Non-members choose emission levels non-cooperatively, while members are supposed to choose emissions such as to maximize the joint welfare of all member countries. Finus and Maus (2008) allow for modest IEAs by specifying that member countries only internalize a fraction of the externalities within the coalition. Given the preference parameter θ_j of the agent who is in charge of the emission choice in country j, agents in member countries set emissions such as to maximize the sum of benefits among all member countries minus a fraction γ of the sum of the agents' damages among all member countries W_i:

$$W_i = \sum_{j \in S} \left[B(e_j) - \gamma \theta_j D(E)\right] , \tag{4}$$

where $S \subseteq I$ denotes the set and $k = |S|$ the number of member countries. The parameter $\gamma \leq 1$ can be interpreted as the level of modesty of a treaty. The case of full internalization, as in the standard coalition formation case, is represented by $\gamma = 1$.

A general criticism against the assumption of member countries maximizing (some fraction of) joint welfare W_i is that it is not in the self-interest of countries to do so. In the case of the standard coalition formation game, i.e., when $\gamma = 1$, this can be rationalized by assuming that member countries set emissions and distribute benefits according to a Nash bargaining solution. Even in this case, one might question why countries behave perfectly non-cooperatively, when they decide about participation, and perfectly cooperatively, once they decided to join the coalition. In fact, member countries individually have an incentive not to comply with maximizing the joint welfare of member countries and to free-ride on the abatement efforts of all other members. In case of modest treaties, i.e., $\gamma < 1$, the explanation of cooperative behavior breaks down, and it is even more unclear why countries should behave as stated in (4).

To circumvent these issues of compliance (or non-compliance, respectively), we present a mechanism in which all countries (i.e., member and non-member countries) make decisions such as to maximize their own welfare given the decisions of all other countries. We show that the outcome of this mechanism is as if member countries behaved according to (4). While the details are relegated to Appendix A.1, the general idea of the mechanism, which is inspired by Gersbach and Winkler (2011), is as follows.

Members in the coalition set up an international emissions permit market, according to the following rules:

1. Participating countries simultaneously and non-cooperatively choose the number of permits they want to issue. Permits can be traded non-discriminatorily across all participating countries (see also Helm, 2003).
2. A fraction μ of issued permits in all participating countries is collected by an international agency (IA) and auctioned directly to firms in member countries.
3. The IA refunds the revenues from auctioned permits to member countries using equal-share lump-sum payments.

Thus, member countries' emission choices are determined in a two stage subgame, in which countries first choose permit issuance, a fraction of which is auctioned by the IA and the revenues are returned lump-sum to member countries. Second, the permit market equilibrium determines the permit price and the emissions in all member countries. We show in the Appendix that there exists a one-to-one correspondence between the fraction μ of permits auctioned by the IA and the degree of modesty γ, where an increasing μ corresponds to an increasing γ. In fact, full auctioning by the IA, $\mu = 1$, corresponds to the standard coalition formation set-up with $\gamma = 1$, while no auctioning via the IA, $\mu = 0$, results in $\gamma = 1/k$, which implies that member countries choose emission levels as non-member countries.[5]

2.3. Weak versus strong delegation

We analyze two different delegation mechanisms, henceforth termed *weak delegation* and *strong delegation*, as coined by Segendorff (1998). The two mechanisms differ in the amount of decision power given to the agent by the principal: While in the weak delegation case the agent's authority is limited to the emission choice, in the strong delegation game the whole decision making process, i.e., both membership and emission choice, is delegated to the agent.

The timing of the *weak delegation* case is as follows:

1. Membership Stage:
   Principals in all countries simultaneously decide whether to join the IEA.
2. Strategic Delegation Stage:
   Principals in all countries simultaneously select an agent.
3. Emission Policy Stage:
   Selected agents in all countries simultaneously choose domestic emissions. Agents in non-member countries choose emissions such as to maximize V_i, while agents in member countries choose emissions such as to maximize W_i.

In the *strong delegation* game, the first two stages are interchanged:

1. Strategic Delegation Stage:
   Principals in all countries simultaneously select an agent.
2. Membership Stage:
   Selected agents in all countries simultaneously decide whether to join the IEA.
3. Emission Policy Stage:
   Selected agents in all countries simultaneously choose domestic emissions. Agents in non-member countries choose emissions such as to maximize V_i, while agents in member countries choose emissions such as to maximize W_i.

Despite being highly stylized, this model captures essential characteristics of the hierarchical structure of domestic and international environmental policy, as we discuss in detail in Section 7.

We solve both games by backward induction. The last stage, the emission policy stage, is structurally identical in both set-ups, as emissions are always chosen by the delegated agents and the membership structure is known at the time of emission choice. In the weak delegation case, we determine in a second step the preferences of the agents, who the principals in member and non-member countries select. Then, we characterize the membership structure for which the international environmental agreement is stable. In the strong delegation case, we first determine the membership structure in equilibrium as a function of the selected agents' preference parameters, before we characterize the principals' optimal choice of agents, which in this case is independent of a country's membership status.

3. Emission policy stage

In the last stage of both the weak and the strong delegation set-up, member and non-member countries are already determined and principals in all countries have delegated the emission policy choice to an agent. Thus, there exists a set $S \subseteq I$ characterizing the $k = |S|$ member countries and a vector $\Theta = (\theta_1, \dots, \theta_n)$ detailing the preference parameters of the selected agents in all countries. Agents in non-member countries $i \notin S$ maximize V_i:

$$\max_{e_i} B(e_i) - \theta_i D(E) \,, \tag{5}$$

[5] Note that $\gamma < 1/k$ would imply that member countries would choose even higher emissions than non-member countries, which would hardly make any economic sense. Thus, this natural lower bound for γ is endogenously derived from our microfoundation.

subject to $E = \sum_{i \in I} e_i$ and given the sum of emissions of all other countries $e_{-i} = E - e_i$. The first-order condition yields the well-known insight that in the Nash equilibrium marginal benefits have to equal marginal environmental damages (from the agent's perspective):

$$B'(e_i) = \theta_i D'(E) \ . \tag{6}$$

Given the sum of emissions of all other countries e_{-i}, the first-order condition (6) implicitly characterizes the best-response function of the agent in country i with respect to the emissions choice e_i.

Analogously, agents in member countries $i \in S$ maximize W_i:

$$\max_{e_i} \sum_{j \in S} \left[B(e_j) - \gamma \theta_j D(E) \right] \ , \tag{7}$$

subject to $E = \sum_{i \in I} e_i$ and given the sum of emissions of all other countries $e_{-i} = E - e_i$. The first-order condition implies that in the Nash equilibrium marginal benefits equal the fraction γ times the sum of damages among all member countries (again, from the perspective of the selected agents):

$$B'(e_i) = \gamma \sum_{j \in S} \theta_j D'(E) \ . \tag{8}$$

Again, the first-order condition implicitly characterizes the agents' best-response functions.

The set of first-order conditions for all non-member and member countries determines the Nash equilibrium with respect to the emission choices in the third stage of the game, which exists and is unique, as the following proposition states:

Proposition 1 (*Unique NE in Emission Policy Stage*)**.** *For any given set S of member countries and any given vector $\Theta = (\theta_1, \dots, \theta_n)$ of preferences of the selected agents, there exists a unique Nash equilibrium of the subgame beginning in stage three, in which the agents of all countries $i \in I$ simultaneously set domestic emission levels e_i such as to maximize either V_i (non-members) or W_i (members), taking the emission choices of all other agents as given.*

As we show in the proof of the proposition,[6] existence follows from the strict concavity of the agents' maximization problem and uniqueness from the curvature properties of the benefit function.

We denote the Nash equilibrium of the subgame beginning in stage three by $\hat{e}(S, \Theta) = \left(\hat{e}_1(S, \Theta), \dots, \hat{e}_n(S, \Theta)\right)$ and the global emission level in this equilibrium by $\hat{E}(S, \Theta)$. For later use, we analyze how the equilibrium emission levels change with a marginal change in the preferences of the selected agent in country i.

Proposition 2 (*Comparative Statics in Emission Policy Stage*)**.** *The following conditions hold for the equilibrium levels of domestic emissions of country $i \in I$, $\hat{e}_i(S, \Theta)$, for the sum of domestic emissions of all other countries $\hat{e}_{-i}(S, \Theta)$ and total emissions $\hat{E}(S, \Theta)$:*

- *For countries $i \notin S$:*

$$\frac{d\hat{e}_i(S, \Theta)}{d\theta_i} < 0 \ , \qquad \frac{d\hat{e}_{-i}(S, \Theta)}{d\theta_i} > 0 \ , \qquad \frac{d\hat{E}(S, \Theta)}{d\theta_i} < 0 \ . \tag{9}$$

- *For countries $i \in S$:*

$$\frac{d\hat{e}_i(S, \Theta)}{d\theta_i} < 0 \ , \qquad \frac{d\hat{e}_{-i}(S, \Theta)}{d\theta_i} \gtreqless 0 \ , \qquad \frac{d\hat{E}(S, \Theta)}{d\theta_i} < 0 \ . \tag{10}$$

Proposition 2 states that domestic emission levels of country i, $\hat{e}_i(S, \Theta)$ (direct effect), always decrease when the preference parameter θ_i increases, i.e., when country i's selected agent cares more about the environment. The sum of emissions of all other countries, $\hat{e}_{-i}(S, \Theta)$ (indirect effect), increases for non-member countries if θ_i increases, due to the strategic substitutability of emission choices. For member countries, the effect is ambiguous. On the one hand, other member countries reduce emissions if θ_i increases, as the sum of marginal damages within the coalition increases. On the other hand, non-member countries increase emissions, due to the strategic substitutability of emission choices. Depending on which effect outweighs the other, $\hat{e}_{-i}(S, \Theta)$ may increase, stay equal or decrease. In any case, the direct effect always outweighs the indirect effect such that global emissions $\hat{E}(S, \Theta)$ are lower in equilibrium when the preference parameter θ_i is higher.

4. Weak delegation

We first analyze the weak delegation set-up, in which principals in the first stage decide whether to join the agreement and in the second stage delegate the emission choice of the third stage to agents.

[6] The proofs of all propositions are relegated to the Appendix.

4.1. Strategic delegation stage

By the logic of backward induction, we first turn to the selection of agents by the principals in the second stage of the game, in which the set S and the number k of member countries is already determined. Formally, the strategic delegation choice of principals is independent of whether the respective country is a member or non-member country. Thus, the principal of country $i \in I$ maximizes:

$$\max_{\theta_i} B\big(\hat{e}_i(S,\Theta)\big) - D\big(\hat{E}(S,\Theta)\big) \ , \tag{11}$$

subject to the equilibrium emissions $\hat{e}_i(S,\Theta)$ and $\hat{E}(S,\Theta)$ of the third stage and given the preference parameter choices θ_j of all other countries $j \neq i$. Then, the first-order condition reads:

$$B'\big(\hat{e}_i(S,\Theta)\big)\frac{d\hat{e}_i(S,\Theta)}{d\theta_i} = D'\big(\hat{E}(S,\Theta)\big)\frac{d\hat{E}(S,\Theta)}{d\theta_i} \ . \tag{12}$$

This equation says that in equilibrium the marginal costs of strategic delegation have to equal its marginal benefits. The costs of choosing an agent with marginally higher environmental preferences (left-hand side) are given by the reduction in domestic benefits, as an agent with higher preference parameter θ_i chooses lower domestic emissions $\hat{e}_i$, while the benefits (right-hand-side) accrue from a reduction in environmental damages due to lower aggregate equilibrium emissions $\hat{E}$.

Inserting the first-order conditions of the third stage, (6) and (8), and the explicit formulae for $d\hat{e}_i/d\theta_i$ and $d\hat{E}/d\theta_i$ for non-member and member countries into Eq. (12), we obtain the following reaction functions for non-member and member countries:

$$\theta_i(\Theta_{-i}) = \frac{1}{1+\phi\left[\sum_{j\notin S, j\neq i}\theta_j + \gamma k\sum_{j\in S}\theta_j\right]} \ , \qquad \forall\ i \notin S \ , \tag{13a}$$

$$\theta_i(\Theta_{-i}) = \frac{k}{\gamma\left[1+\phi\sum_{j\notin S}\theta_j\right]} - \sum_{j\in S, j\neq i}\theta_j \ , \qquad \forall\ i \in S \ , \tag{13b}$$

where Θ_{-i} denotes the vector of preference parameters of all agents but agent i and $\phi = -D''/B'' > 0$.[7]

Bringing $\sum_{j\in S, j\neq i}\theta_j$ to the left-hand side of Eq. (13b), we observe that the Nash equilibrium only depends on the sum of preference parameters over all member countries, which we denote by $\theta^S = \sum_{j\in S}\theta_j$. The economic intuition is that emission choices of the member countries only depend on the sum of marginal damages weighted by the modesty parameter γ, which is given by $\gamma\theta^S D'\big(\hat{E}(S,\Theta)\big)$. In addition, we show in the proof of Proposition 3 that the reaction functions (13a) imply that in equilibrium the principals of all non-member countries choose identical preference parameters for agents, which we denote by θ_i^{NS}. Then, we can re-write Eqs. (13), which determine the choice of preference parameters of the subgame perfect Nash equilibrium starting in the second stage of the weak delegation game, to yield:

$$\theta_i^{NS} = \frac{1}{1+\phi\left[(n-k-1)\theta_i^{NS} + k\gamma\theta^S\right]} \ , \tag{14a}$$

$$\gamma\theta^S = \frac{k}{1+\phi(n-k)\theta_i^{NS}} \ . \tag{14b}$$

In fact, there exists a unique Nash equilibrium for the game starting in the second stage, as the following proposition states:

Proposition 3 (*Unique NE in Strategic Delegation Stage (WD)*)**.** *For any given set S of member countries, there exists a subgame perfect Nash equilibrium of the subgame beginning in stage two, in which principals of all countries $i \in I$ simultaneously select agents such as to maximize U_i taking the choices of all other principals as given. The subgame perfect Nash equilibrium is unique with respect to the preference parameters θ_i^{NS} and θ^S.*

Note that the uniqueness of the Nash equilibrium of the second stage refers to the choice variables θ_i^{NS} and θ^S. In fact, there is a continuum of Nash equilibria in the individual parameter choices θ_i of the principals in member countries $i \in S$, as any combination of θ_i ($i \in S$) satisfying $\sum_{i\in S}\theta_i = \theta^S$ is a Nash equilibrium. However, all these Nash equilibria result in identical emission choices in the third stage and also lead to identical coalition sizes k in the first stage.

In the following, we analyze the properties of the second stage Nash equilibrium. The first important insight is that strategic delegation renders the degree of modesty γ irrelevant. This can be directly seen from Eqs. (13), where both parameters γ and θ^S only show up as the product $\gamma\theta^S$. For any given values of all exogenously given parameters but γ, a change in the degree of modesty γ will – in equilibrium – result in an according change of θ^S such that the product $\gamma\theta^S$ remains unchanged. As also the emission choices in the third stage only depend on the product $\gamma\theta^S$ (see Eq. (8)), emission choices will also be independent of the value of the parameter γ. The decisions about membership in the first stage depend on the anticipated emission levels of the third stage. If these emission levels in equilibrium do not depend on γ, then also the membership decision in the first stage is independent of the modesty parameter γ. This insight is summarized in the following proposition.

[7] Note that both the benefit function B and the environmental damage function D are supposed to be quadratic functions. As a consequence, ϕ is a scalar and does not depend on domestic or global emission levels.

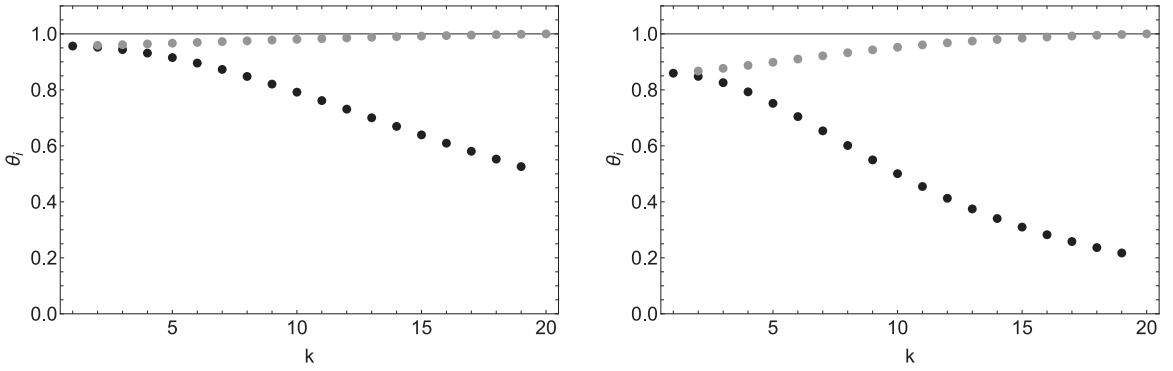

Fig. 1. Illustration of θ_i^{NS} (black dots) and θ_i^S (gray dots) as a function of the stable coalition size k for $n = 20$ and $\phi = 0.025$ (left) and $\phi = 0.01$ (right). For $k = 1$, θ_i^{NS} and θ_i^S coincide. For increasing k, θ_i^S increases while θ_i^{NS} decreases. For the grand coalition $k = n$, $\theta_i^S = 1$ while θ_i^{NS} does not exist.

Proposition 4 (*Modest IEAs are not an Option*)**.** *In the Nash equilibrium of the second stage $\gamma\theta^S$ and θ_i^{NS} do not depend on the parameter γ. As a consequence, neither the emission choices in the emission policy stage nor the participation in the first stage is influenced by the fraction γ.*

Proposition 4 has important consequences for the design of IEAs. As Finus and Maus (2008) show in a setting without strategic delegation, modest IEAs, i.e., $\gamma < 1$, may achieve more than agreements that fully internalize all damages among member countries, as the increase in emissions of every member country due to a decrease in γ may be outweighed by the increase in the stable coalition size. That is, agreements in which each member emits more but the number of members is larger may in sum still emit less than agreements with fewer members but each member emits less. Our analysis shows that increasing the number of members by more modest agreements is not an option in our strategic delegation setting: Principals in member countries completely crowd out any effect of a decreasing γ by delegating to agents with proportionally higher environmental preferences θ_i such that $\gamma\theta^S$ stays constant.

Proposition 4 allows us to drop γ in the weak delegation set-up without loss of generality. Thus, for the remainder of our analysis of the weak delegation framework, we set $\gamma = 1$. In addition, we define $\theta_i^S = \theta^S/k$ as the average preference parameter that principals of member countries choose in equilibrium. Then, the following properties hold for the Nash equilibrium of the game starting in the second stage.

Proposition 5 (*Properties of NE in Strategic Delegation Stage (WD)*)**.** *For the equilibrium preference choices θ_i^{NS} and θ^S of the principals in all countries $i \in I$ the following statements hold:*

- *(i) The equilibrium choices θ_i^{NS} and θ^S do not depend on the set S but only on the number k of member countries. As a consequence, also emission levels in the Nash equilibrium of the third stage only depend on the number k of member countries.*
- *(ii) Equilibrium preference choices θ_i^{NS} of principals in non-member countries decrease with the number k of member countries.*
- *(iii) Average equilibrium preference choices θ_i^S of principals in member countries increase with the number k of member countries.*
- *(iv) For $k = 1$, principals in member and non-member countries delegate to agents with identical preference parameters, i.e., $\theta_i^{NS} = \theta_i^S = \theta^S$. For $k = n$ principals in member countries choose on average agents with the same preferences as they exhibit themselves, i.e., $\theta_i^S = 1$.*
- *(v) For any $1 < k < n$ it holds that $0 < \theta_i^{NS} < \theta_i^S < 1$.*

Proposition 5 part (v) states that principals in both member and non-member countries delegate to agents who evaluate the environmental damages lower than they do themselves (in case of member countries at least on average, as only the sum of preference parameters is uniquely determined). However, principals in non-member countries delegate more strategically than principals in member countries. With an increasing number of member countries k, principals in member countries choose agents with higher preference parameters (part (iii)), and vice versa in non-member countries (part (ii)). Thus, the gap in the preference parameter between agents in member and non-member countries is increasing in k, as shown in Fig. 1. The reason for this result lies in the strategic substitutability of emission choices between different non-member countries and between non-member countries and the coalition of member countries. By delegating to an agent with low environmental preferences, the principal in a non-member country commits to high emissions, which results in decreasing emissions of all other countries. This roll-over of abatement burden to other countries is more attractive, the more the other countries abate and, thus, increases in the coalition size k. Member countries, on the other hand, have to fear less free-riding (at least in absolute terms) the larger is the coalition and, thus, the lower is the number of free-riders. As a consequence, the incentive to delegate to agents with low environmental preferences decreases with coalition size k.

As equilibrium preference parameters only depend on the number of member countries, we denote the Nash equilibrium of the second stage of the game by $\hat{\theta}_i^{NS}(k)$ and $\hat{\theta}^S(k)$. It directly follows that also the emission levels chosen by the agents in the third

stage of the game only depend on the number and not the set of member countries. In addition, the third stage Nash equilibrium is symmetric in the sense that principals of non-member countries select identical agents and principals in member countries only care about the sum of the preference parameters among all the agents in member countries. As a consequence, the emission choices of the agents in the third stage is identical for all agents in non-member countries and identical for all agents in member countries. Thus, by inserting the second stage Nash equilibrium back into the third stage equilibrium emission levels, we obtain:

$$\hat{e}_i^{NS}(k) = \hat{e}_i\big(S, (\hat{\theta}_i^{NS}(k), \hat{\theta}^S(k))\big) , \qquad \forall\, i \notin S , \tag{15a}$$

$$\hat{e}_i^{S}(k) = \hat{e}_i\big(S, (\hat{\theta}_i^{NS}(k), \hat{\theta}^S(k))\big) , \qquad \forall\, i \in S , \tag{15b}$$

$$\hat{E}(k) = (n-k)\hat{e}_i^{NS}(k) + k\hat{e}_i^{S}(k) . \tag{15c}$$

The following proposition states how the equilibrium emission levels change with the number of member countries k.

Proposition 6 (*Equilibrium Emission Levels*)**.** *The following conditions hold for the equilibrium emission levels:*

$$\frac{d\hat{e}_i^{NS}(k)}{dk} > 0 , \quad \frac{d\hat{e}_i^{S}(k)}{dk} \gtreqless 0 , \quad \frac{d\hat{E}(k)}{dk} \gtreqless 0 . \tag{16}$$

Thus, with an increasing number k of member countries equilibrium domestic emission levels of non-member countries increase, while domestic emissions levels of member countries may increase or decrease. The reason why domestic emissions may increase is, again, due to the strategic substitutability of emission choices. While each non-member country emits more if k increases, total emissions of non-member countries may decrease, since there are less non-member countries as the number k of member countries increases. If this is the case, the emissions of member countries are determined by two opposing effects. On the one hand, member countries delegate to agents with higher environmental preferences, which – ceteris paribus – reduces the emissions of member countries. On the other hand, if the sum of emissions in non-member countries is decreasing, this leads – ceteris paribus – to increasing emissions of member countries, due to the strategic substitutability of emission choices. Depending on which effect outweighs the other, domestic emissions in member countries increase or decrease (or may stay the same). As a consequence, also global emissions may increase or decrease in equilibrium with the number k of member countries.

4.2. Membership stage

We now turn to the first stage of the game, in which principals in all countries decide on whether to join the agreement. As usual in the coalition formation literature, the equilibrium number of member countries follows from the conditions of internal and external stability. Therefore, principals evaluate their utility depending on whether or not they are joining the coalition. To this end, we define the utility of principals in member and non-member countries depending on the number of member countries k as:

$$\hat{U}_i^{NS}(k) = \hat{U}_i^{NS}\big(k, \hat{\theta}_i^{NS}(k), \hat{\theta}^S(k)\big) = B\big(\hat{e}_i^{NS}(k)\big) - D\big(\hat{E}(k)\big) , \tag{17a}$$

$$\hat{U}_i^{S}(k) = \hat{U}_i^{S}\big(k, \hat{\theta}_i^{NS}(k), \hat{\theta}^S(k)\big) = B\big(\hat{e}_i^{S}(k)\big) - D\big(\hat{E}(k)\big) . \tag{17b}$$

Then, a coalition is *internally stable* if no principal in a member country would rather leave the coalition, i.e., $\hat{U}_i^S(k) \geq \hat{U}_i^{NS}(k-1)$, and *externally stable* if no principal of a non-member country would rather become a member, i.e., $\hat{U}_i^{NS}(k) > \hat{U}_i^S(k+1)$. Following Hoel and Schneider (1997), we define the stability function as:

$$Z(k) = \hat{U}_i^S(k) - \hat{U}_i^{NS}(k-1) . \tag{18}$$

Then, the equilibrium number $\hat{k}$ of member countries is given by the largest integer for which $Z(k) \geq 0$.[8]

It is well known that even without strategic delegation, no closed form analytical solution for the stable coalition size k can be derived for general bi-quadratic utility functions. As a consequence, we employ the functional forms as specified in (1). Thus, the parameter $\phi = -D''/B''$, as introduced in Section 4.1, equals $\phi = \delta/\beta$. As in equilibrium both the delegation choice in the second and the emission choice in the third stage only depend on ϕ, we can w.l.o.g. set $\beta = 1$ implying $\phi = \delta$. In addition, and again w.l.o.g., we can normalize $\epsilon = 1$, which implies that we measure emissions in fractions of business-as-usual emissions ϵ. Thus, apart from the number of countries n, the model comprises of only one free parameter ϕ.

For this (standard) model specification, the following proposition holds:

Proposition 7 (*Stable Coalition Size in Membership Stage (WD)*)**.** *For the quadratic benefit and damage functions specified in* (1)*, the weak delegation game exhibits a stable coalition size of at most* $\hat{k} = 2$.

While it is cumbersome to formally prove the result of Proposition 7, as shown in the Appendix, the economic intuition is straightforward. Without delegation, the maximum stable coalition size for our welfare specification is well known to be at most two. With strategic delegation, we know from Proposition 5 that non-member countries delegate to less environmentally concerned agents than member countries. As a consequence, member countries abate more relative to non-member countries under strategic delegation compared to the case without delegation. This increases the free-riding incentives for all coalition sizes and, thus, weakly

[8] Note that we employ the usual assumption that countries join the coalition if they are indifferent. This is inconsequential for our results.

decreases the stable coalition size. Therefore, it should come at no surprise that the weak delegation game results in similarly bleak prospects for international environmental cooperation as the standard coalition formation game. In fact, even bleaker, because the possibility to increase stable coalition sizes by lowering the modesty parameter γ is not an option in the weak delegation game, as shown in Proposition 4.

5. Strong delegation

We now turn to the case of strong delegation, in which the principals delegate both the membership decision and the emission choice to the agents. Thus, in the first stage, the principals decide on the agents to whom they delegate. In the second stage, the agents decide whether to join the agreement.

Again, the Nash equilibrium of the membership stage can only be characterized using the functional forms (1) for the benefit and environmental damage function. We employ the same normalization (which, again is w.l.o.g.) as in Section 4.2, i.e., $\epsilon = \beta = 1$ implying $\phi = \delta$. In the strong delegation set-up, delegation cannot be conditioned on membership status, as principals choose agents before membership status is decided by these chosen agents. As a consequence, principals in all countries will choose to delegate to agents with the same preference parameter θ.[9]

As a consequence, emission choices in the third stage are a function of the preference parameter θ, to which principals delegate in the first stage, and the membership structure and, in particular, the number of member countries k, on which agents decide in the second stage. Thus, equilibrium emissions in third stage are given by:

$$\hat{e}_i^{NS}(\theta,k) = 1 - \frac{\phi n\theta}{1+\phi\left[(n-k)\theta+\gamma k^2\theta\right]}\ , \qquad \forall\ i \notin S\ , \tag{19a}$$

$$\hat{e}_i^{S}(\theta,k) = 1 - \frac{\gamma\phi k n\theta}{1+\phi\left[(n-k)\theta+\gamma k^2\theta\right]}\ , \qquad \forall\ i \in S\ , \tag{19b}$$

$$\hat{E}(\theta,k) = \frac{n}{1+\phi\left[(n-k)\theta+\gamma k^2\theta\right]}\ . \tag{19c}$$

5.1. Membership stage

In the second stage of the game, agents of all countries simultaneously decide on whether to join the IEA. Again, we employ the concept of internal and external stability to determine the stable coalition size and define the stability function:

$$Z(k,\theta) = B\big(\hat{e}_i^{S}(\theta,k)\big) - D\big(\hat{E}(\theta,k)\big) - B\big(\hat{e}_i^{NS}(\theta,k-1)\big) + D\big(\hat{E}(\theta,k-1)\big)\ . \tag{20}$$

As in Section 4.2, the stable coalition size $\hat{k}$ is determined by the largest integer for which $Z(\hat{k},\theta) \geq 0$ holds. In the Appendix, we prove the following proposition:

Proposition 8 (*Stable Coalition Size in Membership Stage (SD)*)**.** *For the quadratic benefit and damage functions specified in (1), the strong delegation game exhibits a unique stable coalition size $\hat{k}$, for which the following properties hold:*

(i) The stable coalition size $\hat{k} \in \{k^{min},\dots,k^{max}\}$. While the lower bound $k^{min}(n,\gamma)$ is a function of n and γ, the upper bound $k^{max}(\gamma)$ only depends on γ.

(ii) For given n and γ, which characterize the range $\{k^{min}(n,\gamma),\dots,k^{max}(\gamma)\}$ of attainable stable coalition sizes, the stable coalition size $\hat{k}$ is determined by the product $\psi = \phi\theta$.

In the Appendix, we show that the stability function $Z(k,\theta)$ has a trivial root at $k = 1$, as in this case the domestic welfares of the only member country and all other free-riding countries are identical. In addition, the stability function is concave in k. As a consequence, the stability function may either exhibit another root $k_0 \leq n$, then the stable coalition size is given by the next smaller integer $\hat{k} = \lfloor k_0 \rfloor$, or not exhibit another root $k_0 \leq n$, in which case the grand coalition $\hat{k} = n$ is the stable coalition size.

We further show that the stability function is a quadratic function in $\psi = \phi\theta$. Interpreting the stability function as function of ψ, we can solve for the unique positive value of ψ for which the stable coalition size is k. This solution $\psi(\hat{k})$ characterizes the maximum value of $\psi = \phi\theta$ that renders a coalition of size $\hat{k}$ stable. Then, $k^{max}(\gamma)$ is determined by the next smaller integer of $\bar{k}$ that renders $\psi = 0$, i.e., $k^{max} = \lfloor \bar{k} \rfloor$ with $\psi(\bar{k}) = 0$. $k^{min}(n,\gamma)$ is determined by the next smaller integer of $\underline{k}$ for which ψ diverges to $+\infty$, i.e., $\psi(k)_{k\to\underline{k},k>\underline{k}} = +\infty$. Fig. 2 shows $\psi(k)$ for two different values of γ.

Note that the graphs in Fig. 2 are independent of the exogenously given parameter ϕ and also independent of the preference parameter θ of the agents, to which the principals delegate to in the first stage. In fact, the range of attainable stable coalition sizes is only determined by n and γ (and, in particular, the maximal attainable coalition size k^{max} only depends on the modesty parameter γ). Which of the attainable coalition sizes is realized in the subgame perfect equilibrium, depends on $\psi = \phi\theta$, as shown in the graph.

[9] We shall confirm this in Section 5.2.

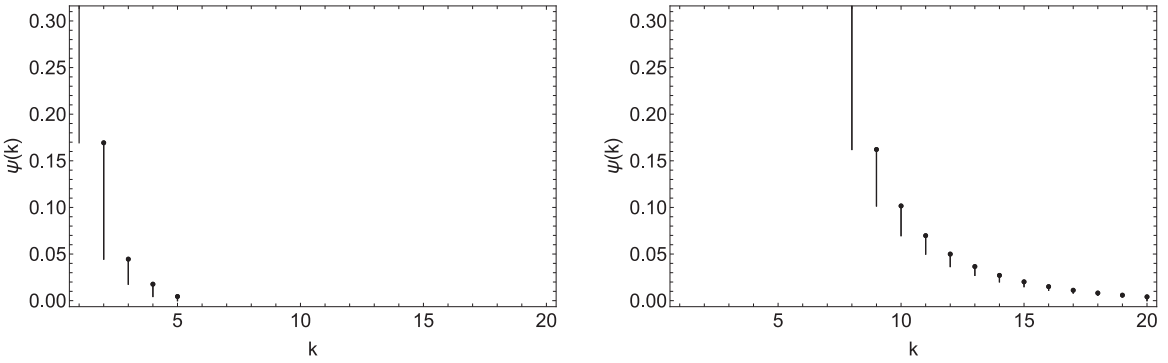

Fig. 2. Illustration of function $\psi(k)$ for $n = 20$ and $\gamma = 0.6$ (left) and $\gamma = 0.15$ (right). The dots indicate the values for $\psi(k)$ obtained by inserting $k \in \{k^{min}, \ldots, k^{max}\}$. The lines indicate the stable coalition sizes $\hat{k}$ for values of ψ between $\psi(k-1)$ and $\psi(k)$. For $\gamma = 0.6$ the attainable values for the stable coalition size range from 1 to 5 (left), while for $\gamma = 0.15$ stable coalition sizes between 6 and 20 are realized depending on $\psi = \phi\theta$.

5.2. Strategic delegation

To determine which of the attainable stable coalition sizes, characterized by the range spanned from k^{min} to k^{max} prevails in the subgame perfect equilibrium of the strong delegation game, we now analyze the first stage. While principals can anticipate the stable coalition size in the subgame perfect equilibrium, as determined by $\psi(k)$ of the second stage of the game, for any coalition size strictly between 1 and n they do not know whether they end up as member or non-member of the coalition. We assume that membership is equally likely for all ex ante identical n countries. Thus, the probability of membership for a given coalition size k is k/n. In addition, we suppose that principals in all countries simultaneously delegate to agents such as to maximize their *expected welfare*:

$$\max_{\theta_i} \quad \frac{k(\Theta)}{n}\left(B\left(e_i^S(\Theta, k(\Theta))\right) - D\left(E(\Theta, k(\Theta))\right)\right) + \frac{n-k(\Theta)}{n}\left(B\left(e_i^{NS}(\Theta, k(\Theta))\right) - D\left(E(\Theta, k(\Theta))\right)\right) . \tag{21}$$

In contrast to the weak delegation case, principals in the strong delegation game cannot condition their choice of agent on whether their own country is a member or non-member of the agreement. This makes an important difference, as the incentives to strategically delegate are different for signatories and non-signatories. As we have seen in Section 4.1, principals of all countries have an incentive to delegate to agents with lower preference parameters than they exhibit themselves, due to the strategic substitutability of emission choices. For principals of member countries, there exists the additional incentive to delegate to an agent with a higher preference parameter than they exhibit themselves in order to counteract the less than full internalization of externalities within the coalition for $\gamma < 1$. We have seen in Proposition 4 that in case of weak delegation any attempt of modesty is fully compensated by the principals delegating to correspondingly "greener" agents and the resulting equilibrium is as if $\gamma = 1$. In the strong delegation game, this full compensation only occurs when the grand coalition is established in the subgame perfect equilibrium, as only in this case principals know for sure that they will become a coalition member.

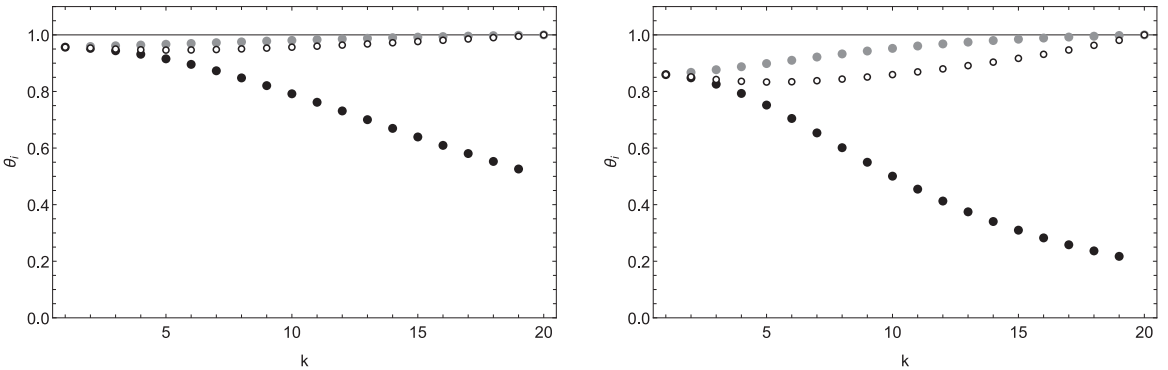

Fig. 3. Illustration of $\hat{\theta}$ (black circles) in the strong delegation game compared to θ_i^{NS} (black dots) and θ_i^S (gray dots) in the weak delegation game as a function of the stable coalition size k for $n = 20$, $\gamma = 1$, and $\phi = 0.025$ (left) and $\phi = 0.01$ (right). For $k = 1$, $\hat{\theta}$, θ_i^{NS} and θ_i^S coincide. For increasing k, $\hat{\theta}$ lies in between θ_i^{NS} and θ_i^S. For the grand coalition $k = n$, $\hat{\theta} = \theta_i^S = 1$, while θ_i^{NS} does not exist.

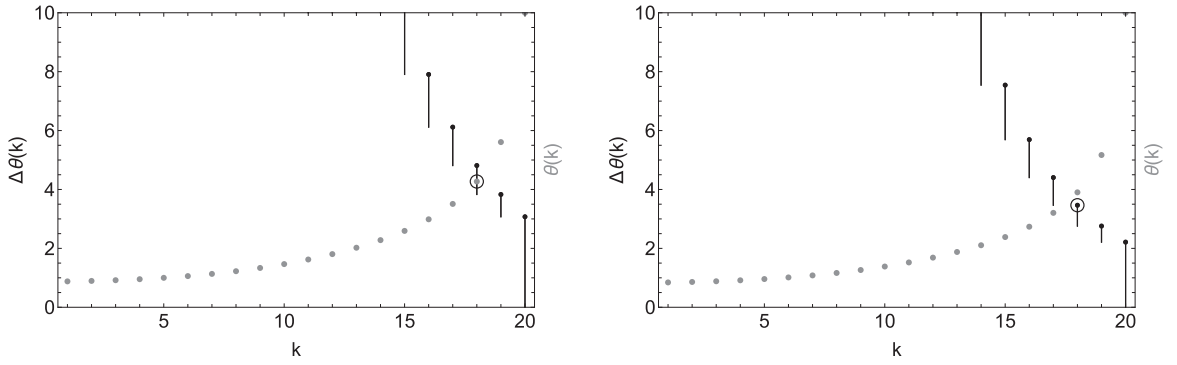

Fig. 4. Illustration of $\Delta\theta(k)$ (black) and $\hat{\theta}(k)$ (gray), which determine the subgame perfect equilibrium (SPE) of the strong delegation game, for $n = 20$ and $\gamma = 0.1$. An "interior"-SPE occurs if $\Delta\theta(k)$ and $\hat{\theta}(k)$ "intersect" (left). If they do not intersect, we obtain a "corner"-SPE (right). The SPE is indicated by the black circle.

The anticipated stable coalition size $k(\Theta)$, which is essentially the inverse of $\psi(k)$, is not a differentiable function, as the stable coalition size k is discrete and jumps from k to $k-1$ whenever θ exceeds $\psi(k)/\phi$. As a consequence, we cannot employ standard differential calculus to derive the principals' best-response functions from maximization problem (21). Instead, we first determine the Nash equilibrium of the first stage of the game for a given stable coalition size k, anticipating the resulting emissions in the third stage given this coalition size k. For a fixed coalition size k, we obtain the following first-order condition for the principal of country $i \in I$:

$$\frac{k}{n} B'\left(e_i^S(\Theta,k)\right)\frac{de_i^S(\Theta,k)}{d\theta_i^S} + \frac{n-k}{n} B'\left(e_i^{NS}(\Theta,k)\right)\frac{de_i^{NS}(\Theta,k)}{d\theta_i^{NS}} = D'\left(E(\Theta,k)\right)\left(\frac{k}{n}\frac{dE(\Theta,k)}{d\theta_i^S} + \frac{n-k}{n}\frac{dE(\Theta,k)}{d\theta_i^{NS}}\right). \tag{22}$$

Eq. (22) is the straightforward generalization of the corresponding first-order condition (12) of the weak delegation game. In equilibrium, the costs of choosing an agent with a marginally higher environmental preference due to a reduction in domestic benefits (left-hand side) have to equal the benefits, which arise through a reduction in environmental damages due to lower aggregate emissions (right-hand side). The difference is that the principal equates expected costs and benefits over the two possibilities of being a signatory or non-signatory country. The following proposition holds:

Proposition 9 (*Unique NE in Strategic Delegation Stage (SD)*)**.** *For the quadratic benefit and damage functions specified in (1) and any given stable coalition size $\hat{k}$, there exists a unique subgame perfect equilibrium (in the sense that principals anticipate emissions in the third stage) of the strong delegation game with the following properties:*

(i) The equilibrium is symmetric, i.e., $\hat{\theta}_i(k) = \hat{\theta}(k)$ for all $i \in I$.
(ii) The equilibrium parameter $\hat{\theta}(k) \leq 1/\gamma$. The equality $\hat{\theta}(n) = 1/\gamma$ only holds in the grand coalition $k = n$.

The choice of preference parameter $\hat{\theta}$ in equilibrium is now perfectly symmetric, as principals cannot condition agent choice on membership status. Note that part (ii) of Proposition 9 is at least on aggregate also true for the preference choice of principals of member countries in the weak delegation case ($\gamma\theta^S \leq k$), as can be seen directly from Eq. (14b). In contrast to the weak delegation set-up, where principals perfectly crowded out any attempt of modest agreements, now all principals choose values for $\hat{\theta}$ which lie in between the corresponding θ_i^{NS} of non-member and θ_i^S of member countries in the weak delegation game (see Fig. 3).

Note that the equilibrium characterized in Proposition 9 is subgame perfect only in the sense that principals correctly anticipate the impact of their preference parameter choices on third stage emissions, but we assume a given and constant stable coalition size k. In fact, the proposition characterizes the set of candidate subgame perfect equilibria of the overall strong delegation game. The remaining step is to match the candidate solutions $\hat{\theta}(k)$ of the first stage with the ranges of θ for which agents implement a particular stable coalition size $\hat{k}$ as determined in the second stage of the game. The range of θ for which agents in the second stage choose a stable coalition size of k is given by:

$$\Delta\theta(k) = \begin{cases} [0, \psi(k^{max})/\phi] , & k = k^{max} \\ (\psi(k+1)/\phi, \psi(k)/\phi] , & k^{min} < k < k^{max} \\ (\psi(k^{min}), \infty) , & k = k^{min} \end{cases}. \tag{23}$$

The subgame perfect equilibria of the strong delegation game are then determined by the "intersection" of $\Delta\theta(k)$ and $\hat{\theta}(k)$. As neither $\Delta\theta(k)$ nor $\hat{\theta}(k)$ is a continuous function, due to the discrete coalition size k, there may not exist an "intersection". In this case, the subgame perfect equilibrium is a corner solution, in which all principals choose $\theta = \psi(k)/\phi$ for the smallest k for which $\hat{\theta}(k)$ exceeds $\psi(k)/\phi$. This is illustrated in Fig. 4. The left panel shows an example of an "interior solution". The set of $\hat{\theta}(k)$ intersects with $\Delta\theta(k)$ for $k = 18$. Thus, if all principals choose $\hat{\theta}(18)$ in the first stage, the stable coalition size determined in the second stage of the game equals $\hat{k} = 18$. Thus, $\hat{\theta}(18)$ is the unique subgame perfect equilibrium of the strong delegation game (indicated by the circle). In the

right panel, we show an example of a "corner" solution. In this case, $\hat{\theta}(k)$ and $\Delta\theta(k)$ do not intersect. Thus, if all principals were to choose $\hat{\theta}(18)$, the agents in the second stage of the game would choose a stable coalition size of $\hat{k} = 17$. For a stable coalition size of 17, the principals would prefer a preference parameter of $\hat{\theta}(17)$, for which the agents would choose a stable coalition size of 18. Thus, neither $\hat{\theta}(18)$ nor $\hat{\theta}(17)$ characterize a subgame perfect equilibrium. In this case the subgame perfect equilibrium is given by $\hat{\theta} = \psi(18)/\phi$, which is the largest preference parameter θ for which agents still choose a stable coalition size of $\hat{k} = 18$ in the second stage (indicated by the circle).[10]

6. Modest IEAs and the grand coalition

The general idea of modest IEAs is to increase the size of participating countries at the expense of coalition members reducing their abatement efforts by internalizing only a fraction γ of the externalities within the coalition. In Section 4.1, we have learned from Proposition 4 that in the weak delegation game any attempt of implementing a modest agreement is perfectly crowded out by the delegation choice of the principals in member countries. As a consequence, the maximum stable coalition size in the subgame perfect equilibrium of the weak delegation game is $\hat{k} = 2$. This implies that the grand coalition can – if at all – be stabilized as a subgame perfect equilibrium if the world only consists of two countries.

The situation is less obvious in the strong delegation game. We have learned from Proposition 8 that there exists a range of attainable coalition sizes, which is determined by the exogenously given parameters γ and n. In particular, k^{max} is given by:

$$k^{max} = \left\lfloor \frac{2 + \sqrt{3 - 2\gamma}}{\gamma} \right\rfloor , \tag{24}$$

and only depends on γ. We directly observe that for γ sufficiently small $k^{max} = n$ can be achieved and, thus, the grand coalition is at least attainable. Moreover, we can show that for γ sufficiently small also $k^{min} \geq n$, as the following proposition states.

Proposition 10 (*Grand Coalition in Strong Delegation Game*)**.** *For a degree of modesty of $\gamma \leq \frac{1}{n-1}$, $k^{min} \geq n$. As a consequence, the grand coalition $k = n$ is the unique subgame perfect equilibrium of the strong delegation game.*

To understand, why the grand coalition can be stabilized in the strong delegation game for sufficiently small γ, recall that the difference to the weak delegation game is twofold: First, the membership decision is taken by the agents. Whenever the agents' preferences differ from the principals', the agents' membership choices in the strong delegation game weakly differ from the principals' membership choices in the weak delegation game. Second, principals cannot condition their choice of agent on whether the own country is a member or non-member of the agreement. The second difference vanishes in case of the grand coalition, as principals correctly anticipate to become a member country for sure. As a consequence, the principals in the strong delegation game choose identical agents as the principals in the weak delegation set-up (see Fig. 3). This implies that also the choice of emissions in the third stage would be the same in both cases if the grand coalition forms. In the strong delegation game, in which the agents decide on membership, the grand coalition is stable because from the agents' point of view the emission choices in the third stage are modest. From the principals' point of view, however, who decide on membership in the weak delegation game, the emission choices of the grand coalition in the third stage are ambitious, as the coalition – in their perspective – fully internalizes all externalities within the coalition. As a consequence, the grand coalition cannot be stable if the world consists of more than two countries.

Finus and Maus (2008) also obtained that the grand coalition can be stabilized for sufficiently small γ. Yet, the stabilization of the grand coalition came at the cost that all coalition members internalize only the fraction γ of externalities and, thus, the resulting equilibrium falls short of the social global optimum (i.e., the outcome that maximizes the sum of welfare over all countries). Intriguingly, this is not the case in the delegation game, as the following Proposition states:

Proposition 11 (*Grand Coalition achieves Global Social Optimum*)**.** *Both in the weak and the strong delegation game, whenever the subgame perfect equilibrium stabilizes the grand coalition, the resulting emission levels are identical to the global social optimum from the principals' point of view.*

The intuition for the result is straightforward. Both in the weak and the strong delegation game, principals in the grand coalition perfectly crowd out γ by delegating to agents with $\theta = 1/\gamma$. Thus, from the principals' point of view, the coalition internalizes all the externalities imposed by any coalition member onto all other coalition members (of course, from the agents' perspective only the fraction γ of externalities is internalized). Yet, in the weak delegation game, the grand coalition can at best be stabilized for $n = 2$ countries, while in the strong delegation game the grand coalition can always be implemented by a sufficiently small choice of γ (see Proposition 10).

This raises the question, who determines the modesty parameter γ? In our analysis, we assumed it to be exogenously given. However, it is straightforward to introduce a zeroth stage, in which principals in all countries decide on γ, for example, by an unanimity vote.[11] Can the principals agree on a value for γ and, if so, on which?

[10] Fig. 4 suggests that $\Delta\theta(k)$ is decreasing, while $\hat{\theta}(k)$ is increasing in k. This being true would constitute a sufficient condition for a unique subgame perfect equilibrium of the strong delegation game. Although we were unable to find any combinations of parameter values for which this does not hold, we were also unable to confirm this conjecture analytically.

[11] The procedure could be governed as follows: a randomly chosen principal suggests a value for γ. If no other principal vetos the proposal it is adopted. Otherwise, another randomly chosen principle may make a suggestion and so on, unless a proposal is adopted.

In the weak delegation game, principals are indifferent between all possible values of γ, as γ is irrelevant for emission levels in the subgame perfect equilibrium and for domestic welfare. Thus, no principal would veto any proposal. In the strong delegation game, principals have an incentive to establish a grand coalition, as this grants them the highest possible welfare, the welfare of – from their perspective – efficient public good provision. Thus, no principal should have an incentive to veto any γ that establishes the grand coalition as the unique subgame perfect equilibrium of the strong delegation game. We see that the strong delegation game together with a preceding agreement on the modesty parameter γ can fully overcome the principals' free-riding incentives and allows them to establish the (from their perspective) first-best outcome.

7. Discussion and conclusions

Both in the weak and the strong delegation game there are two different motives to strategically delegate, i.e., to delegate to agents who have different preferences than the principals themselves. First, principals of all countries have an incentive to delegate to agents exhibiting a lower preference parameter than their own, $\theta < 1$, due to the strategic substitutability of emission choices. By choosing an agent with lower evaluation for the environmental damage, the principal can commit her country to high emission levels, to which the best response of all other countries is to – ceteris paribus – reduce their emission levels. This strategic delegation motive is well understood in the literature on environmental policy and strategic delegation (e.g., Siqueira, 2003; Buchholz et al., 2005; Roelfsema, 2007; Hattori, 2010).

Second, for $\gamma < 1$ principals of member countries have an incentive to delegate to agents that exhibit a higher preference parameter than their own, $\theta > 1$, in order to increase – from the principals' point of view – the "effective" fraction of externalities imposed by the other member countries that they internalize.[12] This incentive is unique to the particular set-up of the coalition formation game and, to the best of our knowledge, has no counterpart in the existing literature on environmental policy and strategic delegation.

As principals of member countries are subject to both strategic delegation motives, their chosen agent may exhibit a preference parameter $\theta \gtreqless 1$, depending on the relative strength of the two. To provide a better intuition for this second strategic delegation incentive and to discuss how it changes between the weak and the strong delegation game, let us suppose that environmental damages are linear in aggregate emission levels, i.e., $D''(E) = 0$. In this case, emission choices in the third stage of the game are governed by dominant strategies and, thus, the first strategic delegation motive vanishes and only the second remains. In the weak delegation game, principals then choose agents with the following preferences:

$$\hat{\theta}_i^{NS} = 1 \,, \qquad \hat{\theta}_i^{S} = \frac{1}{\gamma} \,, \tag{25}$$

independently of the other exogenous parameters n and ϕ. Thus, principals in non-member countries choose "self-representation", i.e., they delegate to agents exhibiting the same preferences as themselves, while principals in member countries perfectly compensate γ by delegating to agents with $\hat{\theta}_i^S = 1/\gamma$. Moreover, for any $n \geq 3$ the stable coalition size is given by $\hat{k} = 3$.

In the strong delegation game, the stable coalition size, as determined in the second stage of the game, is given by:

$$\hat{k} = \min \left[n, \left\lfloor \frac{2 + \sqrt{3 - 2\gamma}}{\gamma} \right\rfloor \right] \,, \tag{26}$$

which is, in particular, independent of the choice of θ in the first stage. The principals in the first stage now choose a $\hat{\theta}$ between $\hat{\theta}_i^{NS} = 1$ and $\hat{\theta}_i^{S} = 1/\gamma$, as they do not know ex ante whether their country will be a member country. In fact, in equilibrium they choose:

$$\hat{\theta}_i = \frac{\gamma \hat{k}^2 + n - \hat{k}}{\gamma^2 \hat{k}^2 + n - \hat{k}} \,, \tag{27}$$

which is also equal to $1/\gamma$ in case of the grand coalition $k = n$, equal to 1 for $k = 1$,[13] and somewhere in between otherwise. The more likely it is that they end up as a coalition member, i.e., the higher is k/n, the closer is their choice of θ to $1/\gamma$, and the higher the chances are to become a non-member, i.e., the higher is $(n - k)/n$, the closer is θ to 1.

Note that the intriguing characteristic of the strong delegation game, i.e., the implementation of the grand coalition for sufficiently small γ and at the same time achieving the first-best from the principals' point of view, survives the simplifying assumption of linear environmental damages. Also the difference in timing between the strong and the weak delegation game is not crucial for this result, as in both cases the principals fully crowd out modesty in the grand coalition. The decisive feature for the result is that in the weak delegation game the principal decides on membership status, while this is the agent's prerogative in the strong delegation game. Intuitively speaking, the principals choose agents who have such a high evaluation for the environmental damage that the agreement from the agents' perspective is so modest that the grand coalition is stable. From the principals' perspective, however, the agreement is strong enough to implement their first-best outcome. While our model analysis is restricted to functional forms that are standard in this literature, there is no reason to believe that our main result crucially hinges on them. In fact, even

[12] Note that in our microfoundation (see Appendix A.1), the emission permit choice of the selected agent translates into a specific degree of internalization in a decentralized manner. In particular, it does not depend on whether the other member countries observe or "recognize" the perceived environmental damages of the agent.

[13] Note that $1 \geq \gamma \geq 1/k$ has to hold and, thus, $\gamma = 1$ for $k = 1$.

biasing our model against strategic delegation as much as possible by assuming linear damages does no harm. As we just argued, it is the particular difference of who decides on membership in the institutional setting that renders weak delegation even worse than no delegation and allows strong delegation to fully overcome the free-riding incentives of global public good provision.

Our model employs two simplifying assumptions that may not survive a reality check:

1. We assume that all countries are identical. While this is clearly an assumption that is not met in reality, we consider it justified, as it allows us to distinguish between inefficiencies stemming from the public good nature of emission abatement and inefficiencies that arise due to countries' heterogeneity. In addition, the introduction of a strategic delegation stage into the two stage coalition formation game stretches the possibility of finding general analytical results to a limit — even for identical countries. Yet, we show in the Appendix for the linear damage specification that all our results can be generalized to a set-up of arbitrarily heterogeneous countries.
2. We assume that principals always find an agent with their preferred preference parameter to which they can delegate. We have seen that in the strong delegation game for sufficiently small γ the grand coalition forms and the principals achieve their first-best emission levels by setting $\theta = 1/\gamma$. Moreover, we have shown that the lower bound for γ to ensure the grand coalition is given by $\gamma = 1/(n-1)$, which corresponds to $\theta = 1/\gamma = n-1$. That is, principals choose agents whose perception of the environmental damage is $n-1$ times as high as their own. Assuming that anthropogenic climate change would be essentially solved if the 10 largest greenhouse gas emitting countries would cooperate (as they account for more than three quarters of global emissions), this would still mean that principals might have to delegate to agents who evaluate climate damages nine times as high as they do themselves. This would be close to a climate change denier delegating to a radical environmental activist. In the Appendix, we investigate how an upper bound of θ impacts on the implementable subgame perfect Nash equilibrium. Assuming 10 countries and that half of the business-as-usual emissions are abated in the principals' first-best outcome, we find that implementing this efficient solution involves delegating to agents, who evaluate climate damages approximately 5 times as high as the principals. In addition, the relationship between the maximal θ and equilibrium abatement levels is concave. For $\theta = 3$ already 83% of the abatement levels and 97% of the welfare levels of the first-best outcome can be achieved.

Another important question is whether and to what extent our highly stylized principal–agent relationship is able to model interactions between domestic and international climate policy. We argue that the timing and the delegation procedure of both the weak and the strong delegation game are compatible with different principal–agent relationships that arise in the hierarchical policy procedures of modern democracies. For example, the principal may be the median voter and the agent an elected government.[14] Then, our weak delegation game translates to a set-up, in which the median voter first decides on the membership status and then elects a government that is in charge of the emission choice. Such a setting could reflect direct democracies, such as Switzerland, where binary and one-shot decisions are often made by the electorate via referendum. In the strong delegation game, the median voter first elects a government, which then decides on membership status and emission levels. Obviously, this might mirror representative democracies, in which the electorate surrenders more decision power to the elected government. Our set-up could also be interpreted as delegation between different levels of government, for example between the legislature and the executive branch. Depending on the political system in a particular country this may rather resemble our weak or strong delegation set-up.

We believe our results have important implications for the future design of IEAs. Unlike most of the existing literature on strategic delegation and environmental policy (e.g., Siqueira, 2003; Buchholz et al., 2005; Habla and Winkler, 2018), we find that strategic delegation is not necessarily an impediment to successful international cooperation. It is less strategic delegation per se but the particular institutional environment in which strategic delegation takes place that determines whether strategic delegation is conducive to overcoming the free-riding incentives of global public good provision. In fact, in both the weak and the strong delegation game, strategic delegation acts as a credible commitment device of the principal to bind herself to a future policy. Whether this commitment ultimately results in better or worse outcomes depends on the incentives imposed by the particular hierarchical governance structure: While in the strong delegation game principals are able to implement their first-best outcome due to strategical delegation, they are even worse off than without delegation in the weak delegation game.

Thus, instead of just analyzing the incentives of existing delegation governance structures, one could also use delegation strategically in the design of international climate policies to overcome the free-riding incentives of public good provision. Obviously, such a governance structure has to be beneficial to all countries, otherwise they would not consent to it. Yet, there is no reason why this cannot be the case. In our model, all principals would willingly adopt the strong delegation framework, as it is in their own best interest. In our opinion, this would constitute a promising avenue for future research.

Acknowledgments

We would like to thank Nadia Ceschi, Thomas Eichner, Wolfgang Habla, Achim Hagen, Bård Harstad, Michael Hoel, Hans Gersbach, Igor Letina, Marc Möller, Robert Schmidt, Alessandro Tavoni, Christian Traeger, Hans-Peter Weikard, the editors, two anonymous reviewers and participants at the WCERE 2018 (Gothenburg), CESifo Area Conference "Energy & Climate Economics"

[14] For this interpretation, we require that $\theta_{i,P} = 1$ is indeed the median in the preference distribution with respect to environmental damages. This can always be achieved by an appropriate normalization. In addition, it is straightforward to show that the voters can be ordered according to the preference parameter θ_i^j, with $\partial \hat{e}_i / \partial \theta_i^j < 0$. As a consequence, the median voter theorem applies.

2018 (Munich), AURÖ meeting 2018 (Münster), EAERE 2020 (Berlin), SURED 2020 (Ascona) and the Annual Meeting of the Verein für Socialpolitik 2021 (Regensburg) and seminar participants at the University of Bern, CESifo Munich, ETH Zurich and European University Viadrina Frankfurt for valuable comments on an earlier draft. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Appendix

A.1. A microfoundation for modest IEAs

In the following, we present a microfoundation for modest IEAs, which rests on the idea of an international permit market with refunding, as developed in Gersbach and Winkler (2011). We refine the emission policy stage of the standard coalition formation set-up by assuming that joining the agreement implies participation in a particular institutional framework. We show that this institutional framework constitutes an incentive compatible mechanism, such that emission abatement, as envisioned by the agreement, is in the best interest of the deciding actors within each country.

All member countries joining the agreement set up an international permit market with refunding, according to the following rules:

1. Participating countries freely choose the number of permits they want to issue. Permits can be traded non-discriminatorily across all participating countries.
2. An (exogenously given) fraction μ of issued permits in all participating countries is collected by an international agency (IA) and auctioned directly to firms in member countries.
3. The IA refunds the revenues for auctioned permits to member countries lump-sum in equal shares.

Thus, the emission policy stage of member countries in both the weak and the strong delegation set-up splits up into two sub-stages:

1. Permit Choice Stage:
   Selected agents in participating countries simultaneously decide on the permit issuance of their country.
2. Permit Market Equilibrium:
   Equilibrium on the international permit market determines the permit price and domestic emissions of all member countries.

In the emission policy stage, member and non-member countries are already determined and principals in all countries have selected an agent. Thus, there exists a set S characterizing the k member countries and a vector $\Theta = (\theta_1, \dots, \theta_n)$ detailing the preference parameters of the selected agents in all countries. We solve the emission policy stage of member countries by backward induction, starting with the permit market equilibrium.

Permit market equilibrium

In the permit market equilibrium, all member countries have already decided on permit issuance. Thus, there exists a vector $\Omega = (\omega_1, \dots, \omega_k)$ detailing the amounts of emission permits issued for all participating countries. We define the total amount of permits by $E^S = \sum_{j \in S} \omega_j$, which also constitutes the supply of permits in the permit market.

The demand for permits (and domestic emissions, respectively) of each member country is derived by maximizing the benefits of domestic emissions minus the costs of permits:

$$\max_{e_i} B(e_i) - p e_i \ , \tag{A.1}$$

which results in the well-known first-order conditions that marginal benefits from emissions have to equal the permit prize p:

$$B'(e_i) = p \ . \tag{A.2}$$

As the marginal benefit function B' is strictly monotonic, the inverse function exists and permit, respectively emission, demand is given by:

$$e_i = B'^{-1}\left[p(E^S)\right] \ . \tag{A.3}$$

As in the permit market equilibrium demand has to equal supply, we obtain:

$$\sum_{i \in S} e_i = \sum_{i \in S} B'^{-1}[p] = E^S \ , \tag{A.4}$$

which constitutes an implicit equation for the equilibrium permit price $p(E^S)$. Inserting back into permit demand yields:

$$e_i(E^S) = B_i'^{-1}\left[p(E^S)\right] \ . \tag{A.5}$$

Permit choice stage

In the permit choice stage, agents of member countries decide on permit issuance ω_i such as to maximize their domestic welfare, anticipating emission choices of member countries determined by the permit market equilibrium and taking emission choices on non-member countries as given. Defining the sum of emissions of non-member countries by $E^{NS} = \sum_{j \notin S} e_i$, the maximization problem of agent $i \in S$ reads:

$$\max_{\omega_i} B\big(e_i(E^S)\big) - \theta_i D(E^S + E^{NS}) + p(E^S)\left[(1-\mu)\omega_i - e_i(E^S)\right] + \frac{\mu}{k} p(E^S) E^S \ . \tag{A.6}$$

Anticipating that $B'(e_i) = p$ in the permit market equilibrium of the last stage, we obtain the following first-order condition:

$$p(E^S)\left[(1-\mu) + \frac{\mu}{k}\right] + p'(E^S)\left[(1-\mu)\omega_i + \frac{\mu}{k} E^S - e_i(E^S)\right] - \theta_i D'(E^S + E^{NS}) = 0 \ . \tag{A.7}$$

Summing up over all member countries $i \in S$ yields:

$$p(E^S)\left[k(1-\mu) + \mu\right] - \sum_{i \in S} \theta_i D'(E) = 0 \ , \tag{A.8}$$

which leads to the equilibrium permit price:

$$p(E^S) = \frac{\sum_{i \in S} \theta_i D'(E)}{k(1-\mu) + \mu} \ . \tag{A.9}$$

Inserting $p(E^S)$ back into $e_i(E^S)$, we obtain:

$$e_i = B_i'^{-1}\left[\frac{\sum_{j \in S} \theta_j D'(E)}{k(1-\mu) + \mu}\right] \ . \tag{A.10}$$

Relationship between μ and ρ

Comparing the emissions of member countries (A.10) with the corresponding emission choice (A.13b), when assuming that member countries maximize some fraction of joint welfare W_i, as given by (4), we find that both are identical if:

$$\gamma = \frac{1}{k(1-\mu) + \mu} \ . \tag{A.11}$$

Thus, there exists a one-to-one correspondence that maps the fraction of permits μ, which is directly auctioned by the international agency and revenues of which are lump-sum refunded to member countries, into the level of modesty γ of an IEA. Whenever μ and k are such that Eq. (A.11) holds, then the permit market refunding mechanism results in emission choices of member countries as if these countries internalized some fraction γ of the emission externalities to all other member countries.

Finally, note that γ is increasing in μ. In fact, we obtain $\gamma = 1/k$ for $\mu = 0$, i.e., without the IA directly auctioning permits, all countries behave as in the non-cooperative emission permit market a la Helm (2003) which is identical to the non-cooperative Nash equilibrium in which all countries simultaneously choose domestic emissions when all countries are identical. For $\mu = 1$, we obtain $\gamma = 1$, i.e., auctioning all permits via the IA, results in emission choices as if all countries took the externalities their emissions impose on all other member countries into account.

A.2. Proof of Proposition 1

(i) Existence:

The maximization problem of country i's selected agent is strictly concave:

$$\mathrm{SOC}_i^{NS} \equiv B_i''(e_i) - \theta_i D_i''(E) < 0 \ , \qquad \forall\ i \notin S \ , \tag{A.12a}$$

$$\mathrm{SOC}_i^{S} \equiv B_i''(e_i) - \gamma \sum_{j \in S} \theta_j D_j''(E) < 0 \ , \qquad \forall\ i \in S \ . \tag{A.12b}$$

Thus, for each country $i \in I$, the reaction function yields a unique best response for any given choices e_j of all other countries $j \neq i$. This guarantees the *existence* of a Nash equilibrium.

(ii) Uniqueness:

Solving the first-order conditions (6) and (8) for e_i, we obtain:

$$e_i = B'^{-1}\left(\theta_i D'(E)\right) \ , \qquad \forall\ i \notin S \ , \tag{A.13a}$$

$$e_i = B'^{-1}\left(\gamma \sum_{j \in S} \theta_j D'(E)\right) \ , \qquad \forall\ i \in S \ . \tag{A.13b}$$

Note that due to assumed curvature properties the marginal benefit function B' is strictly and monotonically decreasing, the inverse functions $B_i'^{-1}$ exist and is also strictly and monotonically decreasing. Summing up emission choices over all countries $i \in I$ yields:

$$E = \sum_{i \notin S} B'^{-1}\left(\theta_i D'(E)\right) + \sum_{i \in S} B'^{-1}\left(\gamma \sum_{j \in S} \theta_j D'(E)\right) \tag{A.14}$$

As the left-hand side is strictly increasing and the right-hand side is decreasing in E, there exists a unique level of total emissions $\hat{E}(S, \Theta)$ in the Nash equilibrium. Substituting back into Eqs. (A.13) yields the unique Nash equilibrium $\hat{e}(S, \Theta)$. □

A.3. Proof of Proposition 2

(i) For country $i \notin S$:
We can write equilibrium emissions $\hat{e}_i(S,\Theta)$ and $\hat{e}_{-i}(S,\Theta)$ as:

$$\hat{e}_i(S,\Theta) = B'^{-1}\left(\theta_i D'\left(\hat{e}_i(S,\Theta)+\hat{e}_{-i}(S,\Theta)\right)\right) , \tag{A.15a}$$

$$\begin{aligned}\hat{e}_{-i}(S,\Theta) \;=\; & \sum_{j\notin S, j\neq i} B'^{-1}\left(\theta_j D'\left(\hat{e}_i(S,\Theta)+\hat{e}_{-i}(S,\Theta)\right)\right) \\ & + \sum_{j\in S} B'^{-1}\left(\gamma \sum_{l\in S}\theta_l D'\left(\hat{e}_i(S,\Theta)+\hat{e}_{-i}(S,\Theta)\right)\right) .\end{aligned} \tag{A.15b}$$

Then, we obtain from the implicit function theorem:

$$\frac{d\hat{e}_i(S,\Theta)}{d\theta_i} = \frac{D'\left(1-\frac{D''}{B''}\left[\sum_{j\notin S, j\neq i}\theta_j+\gamma k\sum_{j\in S}\theta_j\right]\right)}{B''-D''\left[\sum_{j\notin S}\theta_j+\gamma k\sum_{j\in S}\theta_j\right]} , \tag{A.16a}$$

$$\frac{d\hat{e}_{-i}(S,\Theta)}{d\theta_i} = \frac{D'\frac{D''}{B''}\left[\sum_{j\notin S, j\neq i}\theta_j+\gamma k\sum_{j\in S}\theta_j\right]}{B''-D''\left[\sum_{j\notin S}\theta_j+\gamma k\sum_{j\in S}\theta_j\right]} , \tag{A.16b}$$

$$\frac{d\hat{E}(S,\Theta)}{d\theta_i} = \frac{d\hat{e}_i(S,\Theta)}{d\theta_i}+\frac{d\hat{e}_{-i}(S,\Theta)}{d\theta_i} = \frac{D'}{B''-D''\left[\sum_{j\notin S}\theta_j+\gamma k\sum_{j\in S}\theta_j\right]} . \tag{A.16c}$$

(ii) For country $i \in S$:
We can write equilibrium emissions $\hat{e}_i(S,\Theta)$ and $\hat{e}_{-i}(S,\Theta)$ as:

$$\hat{e}_i(S,\Theta) = B'^{-1}\left(\gamma\sum_{j\in S}\theta_j D'\left(\hat{e}_i(S,\Theta)+\hat{e}_{-i}(S,\Theta)\right)\right) , \tag{A.17a}$$

$$\begin{aligned}\hat{e}_{-i}(S,\Theta) \;=\; & \sum_{j\notin S} B'^{-1}\left(\theta_j D'\left(\hat{e}_i(S,\Theta)+\hat{e}_{-i}(S,\Theta)\right)\right) \\ & + \sum_{j\in S, j\neq i} B'^{-1}\left(\gamma\sum_{l\in S}\theta_l D'\left(\hat{e}_i(S,\Theta)+\hat{e}_{-i}(S,\Theta)\right)\right) .\end{aligned} \tag{A.17b}$$

Then, we obtain from the implicit function theorem:

$$\frac{d\hat{e}_i(S,\Theta)}{d\theta_i} = \frac{\gamma D'\left(1-\frac{D''}{B''}\sum_{j\notin S}\theta_j\right)}{B''-D''\left[\sum_{j\notin S}\theta_j+\gamma k\sum_{j\in S}\theta_j\right]} , \tag{A.18a}$$

$$\frac{d\hat{e}_{-i}(S,\Theta)}{d\theta_i} = \frac{\gamma D'\left[(k-1)+\frac{D''}{B''}\sum_{j\notin S}\theta_j\right]}{B''-D''\left[\sum_{j\notin S}\theta_j+\gamma k\sum_{j\in S}\theta_j\right]} , \tag{A.18b}$$

$$\frac{d\hat{E}(S,\Theta)}{d\theta_i} = \frac{d\hat{e}_i(S,\Theta)}{d\theta_i}+\frac{d\hat{e}_{-i}(S,\Theta)}{d\theta_i} = \frac{\gamma k D'}{B''-D''\left[\sum_{j\notin S}\theta_j+\gamma k\sum_{j\in S}\theta_j\right]} . \tag{A.18c}$$

□

A.4. Proof of Proposition 3

(i) We prove by contradiction that the preference parameters of all principals in non-member countries are identical. Therefore, suppose there exists a Nash equilibrium in which $\theta_l \neq \theta_m$ with $l, m \notin S$. Introducing the abbreviation:

$$z = \sum_{j\notin S, j\neq l,m}\theta_j+\gamma k\sum_{j\in S}\theta_j , \tag{A.19}$$

we can write the reaction functions for principle l and m as:

$$\theta_l = \frac{1}{1+\phi(z+\theta_m)} , \qquad \theta_m = \frac{1}{1+\phi(z+\theta_l)} . \tag{A.20}$$

This implies that the following equation has to hold:

$$\theta_l(1+\phi z) = \theta_m(1+\phi z) . \tag{A.21}$$

Obviously, this can only be true if $\theta_l = \theta_m$, which contradicts the assumption of a non-symmetric Nash equilibrium. As a consequence, the Nash equilibrium is given by the system of two Eqs. (14). We also directly observe from (14a) that $\theta_i^{NS} \in (0,1)$, i.e., principals in non-member countries delegate to agents who evaluate the environmental damage lower than they do themselves.

(ii) We prove the existence of a unique equilibrium by showing that the reaction functions intersect exactly once, which determines the preference parameters in the Nash equilibrium. Therefore, we re-write the reaction functions (14) in terms of θ_i^{NS} and the average preference parameter of the coalition $\theta_i^S = \theta^S/k$:

$$\gamma\frac{\theta^S}{k} = \frac{1-\theta_i^{NS}[1+\phi(n-k-1)\theta_i^{NS}]}{\phi k^2\theta_i^{NS}} \equiv R_1(\theta_i^{NS}) \,, \tag{A.22a}$$

$$\gamma\frac{\theta^S}{k} = \frac{1}{1+\phi(n-k)\theta_i^{NS}} \equiv R_2(\theta_i^{NS}) \,. \tag{A.22b}$$

As $\theta_i^{NS} \in (0,1)$, we only have to account for intersections of the two reaction functions in this interval. The following holds:

$$\lim_{\theta_i^{NS}\to 0} R_1(\theta_i^{NS}) = +\infty \,, \qquad R_2(0) = 1 \,, \tag{A.23a}$$

$$R_1(1) = -\frac{n-k-1}{k} < 0 \,, \qquad R_2(1) = \frac{1}{1+\phi(n-k)} > 0 \,. \tag{A.23b}$$

In addition, both reaction functions are strictly monotonically decreasing and strictly convex:

$$R_1'(\theta_i^{NS}) = -\frac{1+\phi(n-k-1)\left(\theta_i^{NS}\right)^2}{\phi k^2\left(\theta_i^{NS}\right)^2} < 0 \,, \tag{A.24a}$$

$$R_1''(\theta_i^{NS}) = \frac{2}{\phi k^2\left(\theta_i^{NS}\right)^3} > 0 \,, \tag{A.24b}$$

$$R_2'(\theta_i^{NS}) = -\frac{\phi(n-k)}{[1+\phi(n-k)\theta_i^{NS}]^2} < 0 \,, \tag{A.24c}$$

$$R_2''(\theta_i^{NS}) = \frac{2\phi^2(n-k)^2}{[1+\phi(n-k)\theta_i^{NS}]^3} > 0 \,. \tag{A.24d}$$

As a consequence, there a exists a unique intersection of R_1 and R_2 on the interval $\theta_i^{NS} \in (0,1)$, which determines the unique Nash equilibrium, for which $\gamma\theta^S \in (k/[1+\phi(n-k)], k)$ holds. This is illustrated in the left panel of Fig. 5. □

A.5. Proof of Proposition 4

We have seen in the proof of Proposition 3 that in the Nash equilibrium of the delegation stage only the product $\gamma\cdot\theta^S$ is uniquely determined. Thus, a ceteris paribus change in γ would in equilibrium result in a corresponding change of θ^S such that the product $\gamma\cdot\theta^S$ remains unchanged. As also the equilibrium emission levels in the third stage only depend on the product $\gamma\cdot\theta^S$, a change in γ would not affect equilibrium emission levels.

For the participation choice in the first stage of the game principals evaluate whether their utility U_i is higher if they become a member of the coalition. As utilities only depend on individual and total emission levels, and these are independent of γ also the participation decision does not depend on γ. □

A.6. Proof of Proposition 5

(i) That the preference parameters in the Nash equilibrium only depend on the number k of member countries and not on their explicit distribution among all n countries follows directly form Eqs. (14) and (A.22).

(ii) and (iii) To show that θ_i^{NS} decreases and θ_i^S increases with k, we first calculate the derivatives of the reaction functions (A.22) with respect to k:

$$\frac{\partial R_1(\theta_i^{NS})}{\partial k} = \frac{\phi(2n-k-2)\left(\theta_i^{NS}\right)^2 - 2(1-\theta_i^{NS})}{\phi k^3\theta_i^{NS}} \,, \tag{A.25a}$$

$$\frac{\partial R_2(\theta_i^{NS})}{\partial k} = \frac{\phi\theta_i^{NS}}{\left[1+\phi(n-k)\theta_i^{NS}\right]^2} > 0 \,. \tag{A.25b}$$

While R_2 is increasing in k for all θ_i^{NS}, R_1 is decreasing if $\theta_i^{NS} < \bar{\theta}$ with:

$$\bar{\theta} = \frac{2}{1+\sqrt{1+2\phi(2n-k-2)}} \,. \tag{A.26}$$

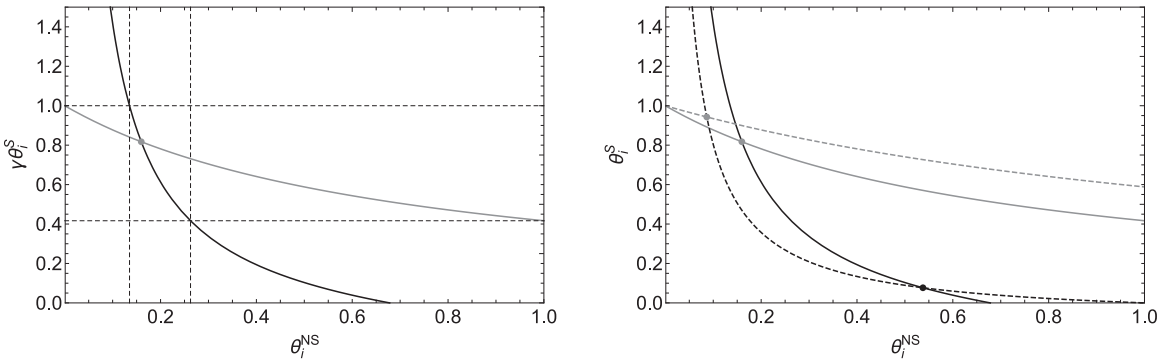

Fig. 5. Illustration of the proofs of Proposition 3 (left) and Proposition 5 (right). The left plot shows the intersection of the reaction functions (R_1 in black and R_2 in gray), which exists and is unique. The horizontal lines show bounds for the feasible range for $\gamma\hat{\theta}_i^S$ in equilibrium, while the vertical lines are bounds for the feasible range of $\hat{\theta}_i^{NS}$ in equilibrium, which are derived by $R_1(\theta_i^{NS}) = 1$ (left bound) and $R_1(\theta_i^{NS}) = 1/(1+\phi(n-k))$ (right bound). The right plot shows how the reaction functions and the resulting equilibrium values change for a increase of k to $k+1$ (solid for k and dashed for $k+1$). While R_2 increases, R_1 tilts anticlockwise around $\bar{\theta}$ (black dot), which implies that R_1 decreases in the range of intersection. As a consequence θ_i^{NS} decreases and $\hat{\theta}_i^S$ increases in equilibrium for an increase in k.

Defining $\Delta R = R_2(\bar{\theta}) - R_1(\bar{\theta})$, we obtain:

$$\Delta R = \frac{(2k-1)\left(1+\sqrt{1+2\phi(2n-k-2)}\right)+2\phi[n(2k-1)-k(k+1)]}{k\sqrt{1+2\phi(2n-k-2)}\left(1+2\phi(n-k)+\sqrt{1+2\phi(2n-k-2)}\right)} > 0\ . \tag{A.27}$$

As $\Delta R > 0$, $\bar{\theta}$ is larger than any feasible equilibrium value $\hat{\theta}_i^{NS}$. As a consequence, $\hat{\theta}_i^{NS}$ decreases and $\hat{\theta}_i^S$ increases when the number of member countries k increases. This is also illustrated in the right panel of Fig. 5.

(iv) For $k = 1$ the coalition is essentially a free-rider to itself, as it consists of only one country, i.e., $V_i = W_i$. Thus, all principals face the same decision problem which results in an identical choice of preference parameter θ_i. For $k = n$, $\theta_i^S = 1$ follows directly from Eq. (A.22b) when setting $\gamma = 1$.

(v) This part follows directly from $0 < \theta_i^{NS} < 1$, as shown in the proof of Proposition 5 and parts (ii), (iii) and (iv) of Proposition 5: For $k = 1$ it holds that $\theta_i^{NS} = \theta_i^S$. For increasing k, θ_i^{NS} is decreasing, while θ_i^S is increasing, ergo $\theta_i^{NS} < \theta_i^S$ and, finally, $\theta_i^S = 1$ for $k = n$. □

A.7. Proof of Proposition 6

From the first-order conditions of the third stage of the game, we obtain:

$$\hat{e}_i^{NS}(k) = B'^{-1}\left(\hat{\theta}_i^{NS} D'[(n-k)\hat{e}_i^{NS}(k) + k\hat{e}_i^S(k)]\right)\ , \tag{A.28a}$$

$$\hat{e}_i^S(k) = B'^{-1}\left(\hat{\theta}^S D'[(n-k)\hat{e}_i^{NS}(k) + k\hat{e}_i^S(k)]\right)\ . \tag{A.28b}$$

Then, the implicit function theorem yields:

$$\frac{d\hat{e}_i^{NS}}{dk} = \frac{\phi\hat{\theta}_i^{NS}(\hat{e}_i^{NS} - \hat{e}_i^S) - \phi\frac{D'}{D''}\left\{[1+\phi k\hat{\theta}^S]\frac{d\hat{\theta}_i^{NS}}{dk} - \phi k\hat{\theta}_i^{NS}\frac{d\hat{\theta}^S}{dk}\right\}}{1+\phi[(n-k)\hat{\theta}_i^{NS} + k\hat{\theta}^S]} > 0\ , \tag{A.29a}$$

$$\frac{d\hat{e}_i^S}{dk} = \frac{\phi\hat{\theta}^S(\hat{e}_i^{NS} - \hat{e}_i^S) + \phi\frac{D'}{D''}\left\{[\phi(n-k)\hat{\theta}^S]\frac{d\hat{\theta}_i^{NS}}{dk} - [1+\phi(n-k)\hat{\theta}_i^{NS}]\frac{d\hat{\theta}^S}{dk}\right\}}{1+\phi[(n-k)\hat{\theta}_i^{NS} + k\hat{\theta}^S]} \gtreqless 0\ . \tag{A.29b}$$

In addition, we know that $\hat{E}(k) = (n-k)\hat{e}_i^{NS} + k\hat{e}_i^S$. Thus:

$$\begin{aligned}\frac{d\hat{E}}{dk} &= (n-k)\frac{d\hat{e}_i^{NS}}{dk} + k\frac{d\hat{e}_i^S}{dk} - \hat{e}_i^{NS} + \hat{e}_i^S \\ &= -\frac{(\hat{e}_i^{NS} - \hat{e}_i^S) + \phi\frac{D'}{D''}\left[(n-k)\frac{d\hat{\theta}_i^{NS}}{dk} + k\frac{d\hat{\theta}^S}{dk}\right]}{1+\phi[(n-k)\hat{\theta}_i^{NS} + k\hat{\theta}^S]} \gtreqless 0\ .\end{aligned} \tag{A.29c}$$

□

A.8. Proof of Proposition 7

We first calculate equilibrium emission levels and domestic welfare for the particular functional forms (1). Setting $\beta = \epsilon = 1$, which is w.l.o.g., as discussed in Section 4.2, we obtain the equilibrium emissions in the third stage as functions of the preference

parameters:

$$\hat{e}_i^{NS}(k,\theta_i^{NS},\theta^S) = 1 - \frac{\phi n \theta_i^{NS}}{1+\phi\left[(n-k)\theta_i^{NS}+k\theta^S\right]}\ , \tag{A.30a}$$

$$\hat{e}_i^{S}(k,\theta_i^{NS},\theta^S) = 1 - \frac{\phi n k \theta_i^{S}}{1+\phi\left[(n-k)\theta_i^{NS}+k\theta^S\right]}\ . \tag{A.30b}$$

$$\hat{E}(k,\theta_i^{NS},\theta^S) = \frac{n}{1+\phi\left[(n-k)\theta_i^{NS}+k\theta^S\right]}\ . \tag{A.30c}$$

Inserting these emission levels yields the following domestic welfares for non-member and member countries:

$$\hat{U}_i^{NS}(k,\theta_i^{NS},\theta_i^S) = \frac{1}{2}\left(1 - \frac{\phi n^2\left[1+\phi\left(\theta_i^{NS}\right)^2\right]}{\left\{1+\phi\left[(n-k)\theta_i^{NS}+k^2\theta_i^S h\right]\right\}^2}\right), \tag{A.31a}$$

$$\hat{U}_i^{S}(k,\theta_i^{NS},\theta_i^S) = \frac{1}{2}\left(1 - \frac{\phi n^2\left[1+\phi k^2\left(\theta_i^{S}\right)^2\right]}{\left\{1+\phi\left[(n-k)\theta_i^{NS}+k^2\theta_i^S\right]\right\}^2}\right). \tag{A.31b}$$

Then, the stability function Z is given by:

$$Z(k) = \hat{U}_i^S\left(k,\hat{\theta}_i^{NS}(k),\hat{\theta}_i^S(k)\right) - \hat{U}_i^{NS}\left(k-1,\hat{\theta}_i^{NS}(k-1),\hat{\theta}_i^S(k-1)\right)\ . \tag{A.32}$$

The stability function (A.32) is difficult to analyze analytically, as it comprises of four different values of the preference parameter θ, for which we cannot derive closed-form solutions to simply plug into the domestic utility functions. As a consequence, we shall analyze the function:

$$\tilde{Z}(k) = \hat{U}_i^S\left(k,\hat{\theta}_i^S(k-1),\hat{\theta}_i^S(k-1)\right) - \hat{U}_i^{NS}\left(k-1,\hat{\theta}_i^S(k-1),\hat{\theta}_i^S(k-1)\right)\ , \tag{A.33}$$

which only includes the preference parameter $\hat{\theta}_i^S(k-1)$. The strategy for proving the proposition is that we first show that $\tilde{Z}(k) > Z(k)$ for all feasible values of $k \in [1,n]$. In a second step, we show that $\tilde{Z}(3) < 0$ holds for all $\phi > 0$ and $n \geq 2$. As $\tilde{Z}(k) > Z(k)$, it holds in particular that $\tilde{Z}(3) > Z(3)$ and, thus, a coalition size of $k = 3$ can never be stable.

(i) $\tilde{Z}(k) > Z(k)$. First, note that the following ordering of the preference parameters holds by virtue of Proposition 5:

$$\theta_i^S(k) > \theta_i^S(k-1) > \theta_i^{NS}(k-1) > \theta_i^{NS}(k)\ . \tag{A.34}$$

Second, we take the derivatives of $\hat{U}_i^S$ with respect to θ_i^{NS} and θ_i^S[15]:

$$\frac{\partial \hat{U}_i^S(k,\theta_i^{NS},\theta_i^S)}{\partial \theta_i^{NS}} = \frac{\phi^2 n^2 (n-k)\left[1+\phi k^2\left(\theta_i^S\right)^2\right]}{\left\{1+\phi\left[(n-k)\theta_i^{NS}+k^2\theta_i^S h\right]\right\}^3} > 0\ , \tag{A.35a}$$

$$\frac{\partial \hat{U}_i^S(k,\theta_i^{NS},\theta_i^S)}{\partial \theta_i^{NS}} = -\frac{\phi n^2 k^2\left[\theta_i^S + \phi(n-k)\theta_i^{NS} - 1\right]}{\left\{1+\phi\left[(n-k)\theta_i^{NS}+k^2\theta_i^S\right]\right\}^3} < 0\ . \tag{A.35b}$$

Thus, $\hat{U}_i^S\left(k,\theta_i^S(k-1),\theta_i^S(k-1)\right) > \hat{U}_i^S\left(k,\theta_i^{NS}(k),\theta_i^S(k)\right)$ as $\theta_i^S(k-1) > \theta_i^{NS}$ and $\partial \hat{U}_i^S/\partial \theta_i^{NS} > 0$, and $\theta_i^S(k-1) < \theta_i^S(k)$ and $\partial \hat{U}_i^S/\partial \theta_i^S < 0$.

In addition, we calculate

$$\begin{aligned}
&\hat{U}_i^{NS}\left(k,\theta_i^S(k),\theta_i^S(k)\right) - \hat{U}_i^{NS}\left(k,\theta_i^{NS}(k),\theta_i^S(k)\right)\\
&\quad= \frac{n^2\phi}{2}\left[\frac{1+\phi\left(\theta_i^{NS}\right)^2}{\left\{1+\phi[(n-k)\theta_i^{NS}+k^2\theta_i^S]\right\}^2} - \frac{1+\phi\left(\theta_i^S\right)^2}{\left\{1+\phi[(n-k)\theta_i^S+k^2\theta_i^S]\right\}^2}\right]\\
&\quad= -\frac{n^2\phi}{2}\left[\frac{\phi\left[1+2\phi k^2\theta_i^{NS}+\phi^2k^2\left(\theta_i^{NS}\right)^2\right]\left[\left(\theta_i^S\right)^2-\left(\theta_i^{NS}\right)^2\right]}{\left\{1+\phi[(n-k)\theta_i^{NS}+k^2\theta_i^S]\right\}^2\left\{1+\phi[(n-k)\theta_i^S+k^2\theta_i^S]\right\}^2}\right.\\
&\qquad\left. + \frac{2\phi^2\theta_i^{NS}\theta_i^S(n-k)\left[1+\phi k^2\theta_i^S\right]\left(\theta_i^S-\theta_i^{NS}\right)}{\left\{1+\phi[(n-k)\theta_i^{NS}+k^2\theta_i^S]\right\}^2\left\{1+\phi[(n-k)\theta_i^S+k^2\theta_i^S]\right\}^2}\right] < 0\ .
\end{aligned} \tag{A.36}$$

Thus, $\tilde{Z}(k) > Z(k)$ for all feasible k, as $\hat{U}_i^S\left(k,\theta_i^S(k-1),\theta_i^S(k-1)\right) > \hat{U}_i^S\left(k,\theta_i^{NS}(k),\theta_i^S(k)\right)$ and $\hat{U}_i^{NS}\left(k-1,\theta_i^S(k-1),\theta_i^S(k-1)\right) < \hat{U}_i^{NS}\left(k-1,\theta_i^{NS}(k-1),\theta_i^S(k-1)\right)$.

[15] Note that $\partial \hat{U}_i^S/\partial \theta_i^{NS} < 0$ holds, because the term in brackets in the numerator can be shown to be positive by substituting $\theta_i^S = 1/[1+\phi(n-k)\theta_i^{NS}]$.

(ii) $\tilde{Z}(3) < 0$. Slightly abusing notation by writing θ instead of $\theta_i^S(k-1)$, we obtain:

$$\begin{aligned}\tilde{Z}(k) &= \hat{U}_i^S(k,\theta,\theta) - \hat{U}_i^S(k,\theta,\theta) \\ &= \frac{n^2\phi}{2}\left[\frac{1+\phi\theta^2}{\{1+\phi[(n-k+1)\theta+(k-1)^2\theta]\}^2} - \frac{1+\phi k^2\theta^2}{\{1+\phi[(n-k)\theta+k^2\theta]\}^2}\right] .\end{aligned} \tag{A.37}$$

The sign of $\tilde{Z}(k)$ is determined by the term in brackets, thus $\tilde{Z}(k) \gtreqless 0$ if and only if $F(k,\theta,\phi) \gtreqless 0$ with:

$$\begin{aligned}F(k,\theta,\phi) &= 4(k-1) + \theta(1-k^2) + \phi\theta[(-4)(n+1) + k(4n+12) - 12k^2 + 4k^3] \\ &+ 2\phi\theta^2[n-k-k^2(n+1)+3k^3-k^4] \\ &+ \phi^2\theta^3[n^2-2nk-k^2(n^2+2n+3)+k^3(6n+10)-k^4(12+2n)+6k^5-k^6] .\end{aligned} \tag{A.38}$$

In addition, we can express ϕ in terms of $\theta_i^S(k-1)$ using Eqs. (A.22):

$$\phi\left(\theta_i^S(k-1)\right) = \phi(\theta) = \frac{(1-\theta)[n-k-(1-\theta)]}{(n-k)\theta^2[(n-k)-k^2(1-\theta)]} \tag{A.39}$$

Note that $\partial\phi(\theta)/\partial\theta < 0$, i.e., θ is the smaller, the larger is ϕ. In fact, $\phi(1) = 0$ and ϕ approaches $+\infty$ for $\theta \to 0$.

Inserting $k = 3$ and $\phi(\theta)$ into F, we obtain:

$$\begin{aligned}F(3,\theta) = \frac{8(1-\theta)}{(n-2)^2(n-6+4\theta)^2}&\left[-144+81n+n^2-4n^3+\theta(396-259n+32n^2+3n^3)\right. \\ &\left.+\ \theta^2(-332+223n-36n^2)+\theta^3(88-57n+9n^2)\right] .\end{aligned} \tag{A.40}$$

Obviously, $F(3,\theta) = 0$ for $\theta = 1$. Note that $\theta = 1$ corresponds to $\phi = 0$. In this case, equilibrium emissions of member and non-member countries are equal to one, i.e., $\hat{e}_i^S = \hat{e}_i^{NS} = 1$, and, thus any coalition size would be stable. For $\theta < 1$, which corresponds to $\phi > 0$, the sign of $F(3,\theta)$ is determined by the terms in brackets, which we denote by $G(\theta)$.

Trying to determine the local extrema of the term in brackets by taking the derivative and setting it equal to zero, we find that for $n \geq 4$, $G(\theta)$ does not exhibit any local extrema and, thus the maximum must be attained at a corner, i.e., at $\theta = 0$ or $\theta = 1$. Evaluation of G at the corners yields:

$$G(0) = -144 + 81n + n^2 - 4n^3 , \tag{A.41a}$$

$$G(1) = -(n-2)^3 < 0 , \tag{A.41b}$$

$$\Delta G = G(1) - G(0) = 152 - 93n + 5n^2 + 3n^3 . \tag{A.41c}$$

As $\Delta G \geq 0$ for all $n \geq 4$ and $G(1) < 0$ this implies that $G(\theta)$ has its maximum at $\theta = 1$ but is negative at the maximum. As a consequence, $G(\theta) < 0$ for all $\theta \in (0,1)$ and $n \geq 4$.

For $n = 3$, we find that $G(\theta)$ is convex and, thus, again exhibits its maximum at the corner. We obtain:

$$G(0)\Big|_{n=3} = 0 , \qquad G(1)\Big|_{n=3} = -1 . \tag{A.42}$$

Thus, also for $n = 3$ we obtain $G(\theta) < 0$ for all $\theta \in (0,1)$. As a consequence, $Z(3) < \tilde{Z}(3) < 0$ always holds and, thus, even a coalition of $k = 3$ can never be stable. □

A.9. Proof of Proposition 8

We first show that there exists a unique stable coalition size. To this end we insert equilibrium emission levels (19) of the third stage into the stability function (20):

$$Z(k,\theta) = \frac{\phi\theta n^2}{2}\left[\frac{1+\phi\theta}{\left[1+\phi\theta(n-k+1+\gamma(k-1)^2)\right]^2} - \frac{1+\phi\gamma^2\theta k^2}{\left[1+\phi\theta(n-k+\gamma k^2)\right]^2}\right] . \tag{A.43}$$

$Z(k,\theta) \lesseqgtr 0$ if and only if $F(n,k,\psi) \lesseqgtr 0$, with

$$F(n,k,\gamma,\psi) = a(n,k,\gamma) + b(n,k,\gamma)\psi + c_1(n,k,\gamma)c_2(n,k,\gamma)\psi^2 , \tag{A.44}$$

where $\psi = \phi\theta$ and

$$a(n,k,\gamma) = -\{1+\gamma[2+k(\gamma k-4)]\} , \tag{A.45a}$$

$$\begin{aligned}b(n,k,\gamma) = -\Big\{1+\gamma\Big[2+\gamma+2n-2\Big(2n+3+2\gamma \\ -k\{2+\gamma[n+4-3k+\gamma(k-1)^2]\}\Big)\Big]\Big\} ,\end{aligned} \tag{A.45b}$$

$$c_1(n,k,\gamma) = n - k\left\{1-\gamma\left[n+1+\gamma(k-1)^2\right]\right\} > 0 , \tag{A.45c}$$

$$c_2(n,k,\gamma) = n - k\left\{1+\gamma\left[n+1-2k+\gamma(k-1)^2\right]\right\} . \tag{A.45d}$$

Z has a root at $k = 1$, as $F(n, 1, 1, \psi) = 0$, which is not surprising, as for $k = 1$ the coalition consists of only one member country, which behaves as the non-member countries. In addition, we show that F is concave in k for $n \geq 3$. The case $k = 2$ is trivial, as either $F(n, 2, \gamma, \psi) \geq 0$, implying the stable coalition size is $\hat{k} = 2$, or $F(n, 2, \gamma, \psi) < 0$ and then the stable coalition size is $\hat{k} = 1$. In either case, there exists a unique stable coalition size.

Taking the second derivative of F with respect to k, we obtain:

$$\frac{\partial^2 F}{\partial k^2} = \frac{\partial^2 a}{\partial k^2} + \underbrace{\frac{\partial^2 b}{\partial k^2}}_{B} \psi + \left(\underbrace{\frac{\partial^2 c_1}{\partial k^2} c_2 + \frac{\partial^2 c_2}{\partial k^2} c_1}_{C} + 2 \frac{\partial c_1}{\partial k} \frac{\partial c_2}{\partial k} \right) \psi^2 , \tag{A.46}$$

with

$$\frac{\partial^2 a}{\partial k^2} = -2\gamma^2 < 0 , \tag{A.47a}$$

$$\frac{\partial^2 b}{\partial k^2} = -4\gamma \left\{ 2 + \gamma \left[6\gamma k(k-1) + n - 9k + 4 + \gamma \right] \right\} , \tag{A.47b}$$

$$\frac{\partial c_1}{\partial k} = \gamma \left[\gamma(3k-1)(k-1) + n + 1 \right] - 1 > 0 , \tag{A.47c}$$

$$\frac{\partial^2 c_1}{\partial k^2} = 2\gamma^2(3k-2) > 0 , \tag{A.47d}$$

$$\frac{\partial c_2}{\partial k} = -\left\{ 1 + \gamma \left[n + 1 + \gamma + 3\gamma k^2 - 4k(1+\gamma) \right] \right\} < 0 , \tag{A.47e}$$

$$\frac{\partial^2 c_2}{\partial k^2} = -2\gamma \left[\gamma(3k-2) - 2 \right] < 0 . \tag{A.47f}$$

Thus, the only terms in (A.46) that are not obviously negative are the terms B and C. We start with C:

$$C = -4\gamma^3 k(3k-2) \left[n - k - 1 + \gamma(k-1)^2 \right] - 4\gamma c_1 < 0 . \tag{A.48}$$

While we cannot show that $B < 0$, we can show that B takes its highest value for $\gamma = 1/k$, as B is decreasing in γ:

$$\frac{\partial B}{\partial \gamma} = -8 \left\{ 1 + \gamma \left[6\gamma k(k-1) + n - 9k + 4 - \gamma \right] \right\} - 4\gamma^2 \left[6k(k-1) + n - 9k + 4 - 1 \right] . \tag{A.49}$$

As $\partial^2 B / \partial \gamma^2 < 0$, $\partial B / \partial \gamma$ is largest for $\gamma = 1/k$. Inserting $\gamma = 1/k$ yields:

$$\left. \frac{\partial B}{\partial \gamma} \right|_{\gamma = \frac{1}{k}} = -8 + \frac{48 - 8n}{k} - \frac{4(n+1)}{k^2} \leq 0 \qquad \forall\, n \geq 3 \wedge n \geq k > 1 . \tag{A.50}$$

Thus, if F is concave for $\gamma = 1/k$ then it is concave for all $\gamma \in [1/k, 1]$. Inserting $\gamma = 1/k$ into $\partial^2 F / \partial k^2$, we obtain:

$$\begin{aligned} \left. \frac{\partial^2 F}{\partial k^2} \right|_{\gamma = \frac{1}{k}} = -\frac{1}{k^4} \Big\{ & 2k^2 + 4k[k(n-k-2)+1]\psi \\ & + 2\left(1 + k\left\{2(n-5) + k[19 + n(n-10) + 2k(2n-5)]\right\}\right)\psi^2 \Big\} . \end{aligned} \tag{A.51}$$

Again, we interpret $\partial^2 F / \partial k^2$ as a function of ψ. For $\psi = 0$, $\partial^2 F / \partial k^2 = -2/k^2 < 0$. Seeking the value ψ for which $\partial^2 F / \partial k^2 = 0$, we obtain:

$$\psi_{1/2} = \frac{-k^2(n-k-2) \pm \sqrt{D}}{1 - 10k + 19k^2 - 10k^3 + 2nk - 10nk^2 + 4nk^3 + n^2k^2} , \tag{A.52}$$

with

$$D = -k^5(n-k) - k^4(5nk - 6n - 14k) - k^3(17k - 6) . \tag{A.53}$$

As $D \leq 0$ for all $n \geq 3$ and $n \geq k > 1$, we obtain that $\partial^2 F / \partial k^2 < 0$. As a consequence, F is concave and can have at most one root in $1 < k \leq n$. If there exists a root k_0 with $1 < k_0 \leq n$, then $\hat{k} = \lfloor k_0 \rfloor$ is the unique stable coalition size. If this root does not exist, then $\hat{k} = n$ if $F > 0$ for $1 < k \leq n$ and $\hat{k} = 1$ if $F < 0$ for $1 < k \leq n$.

Second, we derive the range of attainable stable coalition sizes. Recall that F is a quadratic function in ψ. Thus, we seek ψ for which $F = 0$ holds. This yields a function $\psi(k)$ which gives the maximum ψ that just renders a coalition of size k stable. We obtain two candidate solutions for $\psi(k)$:

$$\psi(k)_{1/2} = \frac{-b \pm \sqrt{b^2 - 4ac_1c_2}}{2c_1c_2} . \tag{A.54}$$

As the lowest possible value of ψ is $\psi = 0$, we derive the maximum coalition size by solving $\psi(k) = 0$ for k:

$$\begin{aligned} \psi(k) = 0 \quad &\Leftrightarrow \quad \pm b = \sqrt{b^2 - 4ac} \quad \Leftrightarrow \quad a = 0 , \\ a = 0 \quad &\Leftrightarrow \quad k = \frac{2 \pm \sqrt{3 - 2\gamma}}{\gamma} . \end{aligned} \tag{A.55}$$

As the lower solution is infeasible, as it would yield $k < 1/\gamma$, the unique solution for the maximum coalition size is given by:

$$k^{max}(\gamma) = \left\lfloor \frac{2+\sqrt{3-2\gamma}}{\gamma} \right\rfloor . \tag{A.56}$$

To determine the minimal stable coalition size, recall that the denominator in Eq. (A.54) reads $2c_1c_2$. While $c_1 > 0$, c_2 is a cubic equation in k, which exhibits one real and two imaginary roots. For the real root $\underline{k} > 1/\gamma$, $c_2 > 0$ for $k < \underline{k}$ and $c_2 < 0$ for $k > \underline{k}$. Thus, $\psi(k)$ diverges for $k \to \underline{k}$. Then $k^{min}(n,\gamma) = \lfloor \underline{k} \rfloor$.

Taking into account that $k \in \left[k^{min}, k^{max}\right]$, $a > 0$ and $c < 0$ in Eq. (A.54). Thus, we obtain:

$$\psi(k) = \frac{-b-\sqrt{b^2-4ac_1c_2}}{2c_1c_2} . \quad \square \tag{A.57}$$

A.10. Proof of Proposition 9

Using the third stage equilibrium emission levels, we obtain the following derivatives:

$$\frac{de_i^{NS}(\Theta,k)}{d\theta_i^{NS}} = -\frac{n\phi\left\{1+\phi\left[\sum_{j\notin S, j\neq i}\theta_j + \gamma k\sum_{j\in S}\theta_j\right]\right\}}{\left[1+\phi\left(\sum_{j\notin S}\theta_j+\gamma k\sum_{j\in S}\theta_j\right)\right]^2} < 0 , \tag{A.58a}$$

$$\frac{de_i^{S}(\Theta,k)}{d\theta_i^{S}} = -\frac{n\gamma\phi\left(1+\phi\sum_{j\notin S, j\neq i}\theta_j\right)}{\left[1+\phi\left(\sum_{j\notin S}\theta_j+\gamma k\sum_{j\in S}\theta_j\right)\right]^2} < 0 , \tag{A.58b}$$

$$\frac{dE(\Theta,k)}{d\theta_i^{NS}} = -\frac{n\phi}{\left[1+\phi\left(\sum_{j\notin S}\theta_j+\gamma k\sum_{j\in S}\theta_j\right)\right]^2} < 0 , \tag{A.58c}$$

$$\frac{dE(\Theta,k)}{d\theta_i^{S}} = -\frac{n\phi\gamma k}{\left[1+\phi\left(\sum_{j\notin S}\theta_j+\gamma k\sum_{j\in S}\theta_j\right)\right]^2} < 0 , \tag{A.58d}$$

Inserting into the first-order condition (22) yields:

$$\begin{aligned} FOC = \frac{n\phi^2}{N^{FOC}}\Bigg[&\gamma k^2 - \gamma^2 k\left(\theta_i + \sum_{j\in S, j\neq i}\theta_j\right)\left(1+\phi\sum_{j\notin S}\theta_j\right) + (n-k) \\ &-(n-k)\theta_i\left\{1+\phi\left[\sum_{j\notin S, j\neq i}\theta_j + \gamma k\sum_{j\in S}\theta_j\right]\right\}\Bigg) , \end{aligned} \tag{A.59}$$

with

$$N^{FOC} = \left[1+\phi\left(\sum_{j\notin S}\theta_j+\gamma k\sum_{j\in S}\theta_j\right)\right]^3 . \tag{A.60}$$

Setting $FOC = 0$ and solving for θ_i, we obtain the reaction function for the principal of country i:

$$\theta_i(\Theta_{-i}) = \frac{(n-k)+\gamma k^2 - \gamma^2 k\sum_{j\in S, j\neq i}\theta_j\left(1+\phi\sum_{j\notin S}\theta_j\right)}{\gamma^2 k\left(1+\phi\sum_{j\notin S}\theta_j\right)+(n-k)\left[1+\phi\left(\sum_{j\notin s, j\neq i}\theta_j+\gamma k\sum_{j\in S}\theta_j\right)\right]} \tag{A.61}$$

First, we show that only symmetric equilibria exist by contradiction. To this end, we define:

$$\begin{aligned} a &= (n-k)+\gamma k^2 , & b &= \gamma^2 k\left(1+\phi\sum_{j\notin S}\theta_j\right) , \\ c &= (n-k)\left(1+\phi\gamma k\sum_{j\in S}\theta_j\right) , & d &= (n-k)\phi , \\ \theta^S &= \sum_{j\in S, j\neq m,l}\theta_j , & \theta^{NS} &= \sum_{j\notin S, j\neq l,m}\theta_j , \end{aligned} \tag{A.62}$$

and assume that $\theta_l \neq \theta_m$ for two countries $l \neq m$. Then the following two conditions have to hold in equilibrium:

$$\theta_l = \frac{a-b\left(\theta_m+\theta^S\right)}{b+c+d\left(\theta_m+\theta^{NS}\right)} , \qquad \theta_m = \frac{a-b\left(\theta_l+\theta^S\right)}{b+c+d\left(\theta_l+\theta^{NS}\right)} . \tag{A.63}$$

This is equivalent to:

$$\left(b+c+d\theta^{NS}\right)\left[a-b\left(\theta^S+\theta_m+\theta_l\right)\right]\left(\theta_l-\theta_m\right) - bd\theta_m\theta_l\left(\theta_l-\theta_m\right) = 0 . \tag{A.64}$$

As $\theta_l \neq \theta_m$, we can divide by $(\theta_l - \theta_m)$ and obtain:

$$\theta_l = \frac{b+c+d\theta^{NS}}{b} \frac{a-b\left(\theta_m+\theta^S\right)}{b+c+d\left(\theta_m+theta^{NS}\right)} , \tag{A.65}$$

which contradicts Eqs. (A.63), as the first fraction is not equal to 1. Thus, equilibria have to be symmetric, i.e., $\theta_l = \theta_m$ for all $l, m \in I$.

For symmetric $\theta = \theta_i$ for all $i \in I$, the first-order condition is zero if and only if the following equation holds:

$$\underbrace{\phi(n-k)\left[(n-k)+\gamma k^2+\gamma^2k^2-1\right]}_{A\geq 0}\theta^2 + \underbrace{\left[(n-k)+\gamma^2k^2\right]}_{B>0}\theta - \underbrace{\left[(n-k)+\gamma k^2\right]}_{C>0} = 0 . \tag{A.66}$$

For $k = n$ this reduces to

$$\gamma^2k^2\theta - \gamma k^2 = 0 , \tag{A.67}$$

the solution of which is $\hat{\theta}(n) = 1/\gamma$. As $A > 0$ for $1 < k < n$, we directly obtain that $\hat{\theta}(k) < \frac{1}{\gamma}$ for $1 < k < n$. The unique solution for $1 < k < n$ is given by:

$$\hat{\theta}(k) = \frac{-B+\sqrt{B^2+4AC}}{2A} . \tag{A.68}$$

It remains to show that the first-order conditions (A.59) characterize the best-response functions of the principals, i.e., we have to show that the second-order conditions hold for our candidate equilibria $\hat{\theta}(k)$. Taking the derivative of the first-order condition (A.59) with respect to θ_i and taking into account that the equilibrium is symmetric, yields for the second-order condition:

$$\begin{aligned} SOC = \frac{n\phi}{N^{SOC}} \Big\{ & 2\gamma^3k^3\phi\theta\left[1+\phi(n-k)\theta\right] - \gamma^2k\left[1+\phi(n-k)\theta\right]^2 - \gamma^2k^3\phi - 2\gamma^2k^3\phi \\ & + 2(n-k)\phi\theta\left\{1+\phi\left[(n-k-1)\theta+\gamma k^2\theta\right]\right\} - 3(n-k)\phi \\ & - (n-k)\left\{1+\phi\left[(n-k-1)\theta+\gamma k^2\theta\right]\right\}^2\Big\} , \end{aligned} \tag{A.69}$$

with

$$N^{SOC} = \left[1+\phi\left(\sum_{j\notin S}\theta_j + \gamma k\sum_{j\in S}\theta_j\right)\right]^4 . \tag{A.70}$$

The second-order condition is satisfied if $SOC < 0$ which holds if and only if the term in curly brackets is negative. Re-arranging this term yields:

$$\begin{aligned} & -\left\{2\gamma^2k^3\phi(1-\gamma\theta) + \phi^2\theta^2\gamma^2(n-k)\left[k^4+(n-k)k-2\gamma k^3\right]\right. \\ & + 2(n-k)\phi\theta\left[(n-k-1)+\gamma k^2-1\right] + \gamma^2k\left[1+2\phi\theta(n-k)\right] + 3\phi\gamma^2k^3 \\ & \left. +(n-k)(1+3\phi) + \phi^2\theta^2(n-k)\left[(n-k-1)+\gamma k^2\right](n-k-3)\right\} . \end{aligned} \tag{A.71}$$

All terms in curly brackets but the last are always positive. The last term is non-negative for $(n-k) \geq 3$ and equal to zero for $n = k$. Thus, the remaining cases we have to check are $k = n-2$ and $k = n-1$. To do so, we concentrate on the terms in the second-order condition containing $\phi^2\theta^2$, since all other terms are negative anyway:

$$\phi^2\theta^2(n-k)\underbrace{\left\{2\gamma^3k^3+(n-k-1)\left[2-(n-k-1)-2\gamma k^2\right]+2\gamma k^2-\gamma^2k^2(n-k)-\gamma^2k^4\right\}}_{\Delta} \tag{A.72}$$

We have to show that $\Delta \leq 0$. For $k = n-2$, we obtain:

$$\Delta = \gamma^2k^3(2\gamma-k) + 1 - 2\gamma^2k^2 . \tag{A.73}$$

Δ is largest for $k = 1$, which also implies $\gamma = 1$ for which $\Delta = 0$. In addition, $\Delta < 0$ for all $k \geq 2$. For $k = n-1$, Δ reduces to:

$$\Delta = \gamma k^2\left[2(\gamma^2k+1)-\gamma(k^2+1)\right] \tag{A.74}$$

It can easily be shown that $\Delta < 0$ for $k \geq 3$. However, for $k < 3$ Δ can be positive. To show that the second-order conditions also hold in these cases, recall that θ is given by:

$$\hat{\theta}(k) = \frac{-B\pm\sqrt{B^2-4AC}}{2A} \leq \frac{-B+\sqrt{B^2}+\sqrt{4AC}}{2A} = \frac{\sqrt{AC}}{A} , \tag{A.75}$$

where A and C yield for $n-k = 1$:

$$A = \phi\gamma k^2(1+\gamma) , \quad C = 1+\gamma k^2 . \tag{A.76}$$

Thus, we obtain as an upper bound for $\phi^2\theta^2$:

$$\phi^2\theta^2 \leq \frac{\phi(1+\gamma k^2)}{\gamma k^2(1+\gamma)} , \tag{A.77}$$

For $k = 1$, also $\gamma = 1$ and thus $\phi^2\theta^2 \leq \phi$. For $k = 2$ it holds that $1/2 < \gamma < 1$ and thus also $\phi^2\theta^2 \leq \phi$ holds. We now collect all terms with $\phi^2\theta^2$ and with ϕ in the second-order condition and use $n - k = 1$ and $\phi^2\theta^2 \leq \phi$ to obtain:

$$\phi\left\{2\gamma k^2(\gamma^2 k + 1) - \gamma^2 k^2(k^2 + 1) - 3\gamma^2 k^3 - 3\right\} \tag{A.78}$$

Inserting $k = 1$ yields:

$$2\gamma^3 + 2\gamma - 5\gamma^2 - 3 = 2\gamma^2(\gamma - 1) - 2(\gamma - 1) - 3\gamma^2 - 1 < 0\ . \tag{A.79}$$

For $k = 2$, we obtain:

$$16\gamma^3 + 8\gamma - 44\gamma^2 - 3 = 16\gamma^2(\gamma - 1) + \gamma(8 - 28\gamma) - 3 < 0\ , \tag{A.80}$$

as $1/2 \leq \gamma \leq 1$. Thus, the second-order conditions hold in all possible cases and the symmetric equilibrium is given by $\hat{\theta}(k)$. □

A.11. Proof of Proposition 10

Recall from Eqs. (A.45):

$$c_2(n, k, \gamma) = n - k\left\{1 + \gamma\left[n + 1 - 2k + \gamma(k - 1)^2\right]\right\}\ , \tag{A.81}$$

the unique real root $\underline{k}$ which determines the minimal attainable stable coalition size k. If $\underline{k} \geq n$, then the unique subgame perfect Nash equilibrium of the strong delegation game is characterized by $\hat{k} = n$, $\hat{\theta}(n) = 1/\gamma$ and the corresponding third stage emission levels. We obtain:

$$c_2(n, n, \gamma) = -n\gamma\left[n - 1 + \gamma(n - 1)^2\right]\ , \tag{A.82}$$

which is equal to zero if and only if:

$$-\gamma^2 n(n - 1)^2 - \gamma n(n - 1) = 0\ . \tag{A.83}$$

This equation holds for $\gamma = 0$ and $\gamma = 1/(n - 1)$. As $\gamma = 0$ is not feasible, the unique solution is given by $\gamma = 1/(n - 1)$. Thus, when $\gamma < 1/(n - 1)$, the grand coalition is the unique subgame perfect Nash equilibrium of the strong delegation game. □

A.12. Proof of Proposition 11

In the grand coalition of both the strong and the weak delegation game, the principals delegate to agents with the preference parameter $\theta = 1/\gamma$. This can be seen from (14b) for $k = n$ for the weak delegation game and from part (ii) of Proposition 9 in case of the strong delegation game. Then, the first-order condition for all agents in the third stage (8) reads:

$$B'(e_i) = \gamma \sum_{j \in S} \theta_j D'(E) = \gamma k \frac{1}{\gamma} D'(E) = k D'(E)\ . \tag{A.84}$$

Obviously, this is the Lindahl–Samuelson condition for efficient public good provision from the principals' perspective. □

A.13. Robustness check: Heterogeneous countries

In our model framework we assume countries to be identical. In the following, we show that our main results, i.e., there is no alternative to "narrow-but-deep" in the weak delegation game and that principals in the strong delegation game can achieve the efficient solution from their point of view, are generalizable for heterogeneous countries. For reasons of brevity, we only consider the linear damage specification. Thus, we consider the following benefit and damage functions:

$$B_i(e_i) = \beta_i e_i\left(\epsilon_i - \frac{1}{2}e_i\right)\ , \quad D_i(E) = \delta_i E\ , \tag{A.85}$$

with country specific exogenous parameters β_i, δ_i and ϵ_i $(i \in I)$.

A question that arises in a setting of heterogeneous countries is how member countries distribute the cooperation gain among themselves. To render the exposition as simple as possible, we assume that an IEA now also specifies, in addition to the degree of modesty γ, country and membership structure specific shares $\tau_i(S)$ for all countries $i \in S$ and all membership structures S such that:

$$\sum_{i \in S} \tau_i(S) = 1\ , \qquad \forall\ i \in S,\ \forall S \in \mathcal{P}(S)\ , \tag{A.86}$$

where $\mathcal{P}(S)$ denotes the power set of S. For our microfoundation outlined in Appendix A.1, this translates into country and membership structure specific refunding shares of the revenues of the international agency. Essentially, the shares $\tau_i(S)$ define a transfer scheme allowing member countries to disentangle efficiency and distribution.

Again, the third stage, i.e., the choice of emissions by the agents, is identical in both the weak and the strong delegation game. The selected agents maximize their respective welfare functions and equilibrium emission choices, determined by the corresponding first-order conditions, are given by:

$$e_i^{NS} = \epsilon_i - \frac{\delta_i}{\beta_i}\theta_i\ , \tag{A.87a}$$

$$e_i^S = \epsilon_i - \frac{\gamma \sum_{j\in S} \delta_j \theta_j}{\beta_i} \ . \tag{A.87b}$$

Consequently, global emissions sum up to:

$$\begin{aligned} E &= \sum_{j\notin S} \left(\epsilon_j - \frac{\delta_j}{\beta_j}\theta_j \right) + \sum_{j\in S} \left(\epsilon_j - \frac{\gamma \sum_{k\in S} \delta_k \theta_k}{\beta_j} \right) \\ &= \sum_{j\in I} \epsilon_j - \sum_{j\notin S} \frac{\delta_j \theta_j}{\beta_j} - \gamma \sum_{j\in S} \delta_j \theta_j \sum_{j\in S} \frac{1}{\beta_j} \end{aligned} \tag{A.88}$$

First, we consider the weak delegation game. In Stage 2, principals in non-member countries $i \notin S$ delegate to their preferred agent θ_i such as to maximize:

$$\max_{\theta_i} \beta_i e_i^{NS} \left(\epsilon_i - \frac{1}{2} e_i^{NS} \right) - \delta_i \left(\sum_{j=1}^{n} \epsilon_j - \sum_{j\notin S} \frac{\delta_j \theta_j}{\beta_j} - \gamma \sum_{j\in S} \delta_j \theta_j \sum_{j\in S} \frac{1}{\beta_j} \right) \ . \tag{A.89}$$

From the corresponding first-order conditions, we find that the optimal agent choice for non-member principals is self-representation, which is analogous to the linear damages case for homogeneous countries:

$$\hat{\theta}_i^{NS} = 1 \ , \qquad \forall\ i \notin S \ . \tag{A.90}$$

For principals in member countries $i \in S$, the optimization problem is given by:

$$\max_{\theta_i} \tau_i(S) \sum_{j\in S} \left(B_j^S(e_j^S(\Theta)) - D_j(E(\Theta)) \right) \ . \tag{A.91}$$

The corresponding first-order condition simplifies to:

$$\sum_{j\in S} \delta_j \theta_j = \frac{1}{\gamma} \sum_{j\in S} \delta_j \ . \tag{A.92}$$

As can easily be seen from Eq. (A.87b), all distributions of θ_i among member countries that satisfy Eq. (A.92) result in the same amount of emissions of member countries. As a consequence, we can concentrate on the symmetric equilibrium $\theta_i^S = 1/\gamma$.

Thus, we find that principals in member countries, like in the case of homogeneous countries, delegate to agent such as to fully crowd out the modesty parameter γ. Consequently, for the membership stage, we find ourselves in the standard specification of the coalition formation game with linear damages and heterogeneous countries, i.e., the membership decision is taken as if $\gamma = 1$ and, thus, modest environmental agreements are not an option (see Proposition 4).

Second, we show for the strong delegation game that principals can implement the first-best from their point of view for appropriately chosen parameters γ and $\tau_i(S)$. In a first step, we suppose that the grand coalition forms and show that under this assumption the optimal delegation choices of principals in the first stage lead to the first-best outcome from the principals' point of view. In a second step, we determine the parameters γ and $\tau_i(S)$ such that the grand coalition is indeed stable.

Given that the grand coalition $S = I$ forms, in the first stage the principal in country i solves:

$$\max_{\theta_i} \tau_i(I) \sum_{i\in I} \left[B_i(e_i^S(\Theta)) - D_i(E(\Theta)) \right] \ , \tag{A.93}$$

yielding the following first-order conditions:

$$\sum_{j\in I} \delta_j \theta_j = \frac{1}{\gamma} \sum_{j\in I} \delta_j \ . \tag{A.94}$$

Inserting Eq. (A.94) into the emission choices of the third stage, we obtain

$$e_i^S = \epsilon_i - \frac{\sum_{j\in I} \delta_j}{\beta_i} \ , \tag{A.95}$$

which is equal to the principals' first-best emission levels. Thus, given that the grand coalition forms, principals can fully overcome the free-riding incentives of public good provision, as in the case of homogeneous countries (see Proposition 11).

In a second step, we show that there exist exogenous parameters γ and $\tau_i(I)$ such that the grand coalition is stable. The stability function of country i is given by:

$$Z_i(S) = \tau_i(S) \sum_{j\in S} W_j(j \in S, S) - W_i(i \notin S, S \setminus i) \ . \tag{A.96}$$

Summing up over all countries $i \in I$, we obtain the following condition, which must hold for the grand coalition to be just stable:

$$\sum_{j\in I} W_j(j \in I, I) = \sum_{j\in I} W_j(j \notin I, I \setminus i) \ . \tag{A.97}$$

This condition says that the welfare in the grand coalition must (at least) equal the gains from unilateral deviation of all countries and is an implicit equation for the unique maximum degree of modesty γ such that grand coalition is stable.[16] Inserting γ into the

[16] Note that the left-hand side does not depend on γ, while the right-hand side is a function of γ^2. Thus, condition (A.97) is a quadratic function of γ with a unique positive root.

Table 1
Optimal values for $\hat{\theta}$ and γ when principals' delegation choices are restricted by an upper bound θ^{max}. In addition, the table shows the corresponding values for $\gamma\hat{\theta}$, abatement levels $\hat{a}$ relative to $a^\star$ and welfare levels $\hat{W}$ relative to $W^\star$.

θ^{max}	$\hat{\theta}$	γ [%]	$\gamma\hat{\theta}$ [%]	$\hat{a}$ [% $a^\star$]	$\hat{W}$ [% $W^\star$]
1	1	29.13	29.13	45.12	69.88
1.5	1.5	27.32	40.98	58.13	82.47
2	2	25.90	51.80	68.25	89.92
2.5	2.5	24.74	61.86	76.44	94.45
3	3	23.78	71.35	83.28	97.20
3.5	3.5	22.96	80.37	89.12	98.82
4	4	22.26	89.02	94.19	99.66
4.5	4.5	21.64	97.36	98.66	99.98
5	4.74	21.09	1	1	1

stability function (A.96) of country i, we obtain for $\tau_i(I)$:

$$\tau_i(I) = \frac{W_i(i \notin i, I \setminus i)}{\sum_{j \in i} W_j(j \in I, I)} \,. \tag{A.98}$$

In combination, the Eqs. (A.97) and (A.98) pin down the modesty parameter and transfer scheme, for which the grand coalition is stable in case of heterogeneous countries.

A.14. Robustness check: Restriction on the upper bound θ^{max}

In our model analysis we have assumed that principals can always delegate to their preferred agent. Essentially, this implies that there is no upper bound for θ^{max} (or if it exists it never binds). We have shown in Propositions 10 and 11 that in the strong delegation game the grand coalition is the unique subgame perfect equilibrium if the degree of modesty γ is sufficiently small and that the principals can achieve the first-best outcome from their point of view by delegating to agents with $\theta = 1/\gamma$. In the following, we analyze how the attainable equilibrium in the strong delegation game depends on θ^{max}.

We assume $n = 10$ countries and $\phi = 0.01$, which implies that in the principals' first-best outcome half of the business-as-usual emissions would be abated. Setting different values for θ^{max}, we first calculate the corresponding γ such that the grand coalition is stable given that principals would delegate to agents with $\theta = \theta^{max}$. Given the grand coalition is stable, principals would prefer to delegate to agents with $\theta^{opt} = 1/\gamma$. However, if $\theta^{opt} > \theta^{max}$, the best the principals can do is to delegate to agents with $\theta = \theta^{max}$. In this case, abatement and welfare levels in the subgame perfect Nash equilibrium will fall short of the corresponding levels in the principals' first-best outcome.

Table 1 shows the results. The principals' first-best outcome is achieved by a combination of $\gamma = 21.09\%$ and $\hat{\theta} = 4.74$. This implies that, if $\theta^{max} < 4.74$, the principals' first-best cannot be implemented as the unique subgame perfect Nash equilibrium of the strong delegation game. However, we observe that the relationship between θ^{max} and abatement level $\hat{a}$ is concave. For $\hat{\theta} = 4$, we already achieve 94.19% of the abatement and 99.66% of the welfare levels of the first-best. These values drop to 83.28% and 97.20%, and 68.25% and 89.92% respectively, if θ^{max} is restricted to 3, respectively 2.

References

Aldy, J.E., Barrett, S., Stavins, R.N., 2003. Thirteen plus one: a comparison of global climate policy architectures. Clim. Policy 3, 373–397.
Barrett, S., 2002. Consensus treaties. J. Inst. Theor. Econ. 158, 529–547.
Barrett, S., 2003. Environment and Statecraft: The Strategy of Environmental Treaty-Making. Oxfrod University Press.
Battaglini, M., Harstad, B., 2020. The political economy of weak treaties. J. Polit. Econ. 128, 544–590.
Besley, T., Coate, S., 2003. Centralized versus decentralized provision of local public goods: a political economy approach. J. Public Econ. 87 (12), 2611–2637.
Buchholz, W., Haupt, A., Peters, W., 2005. International environmental agreements and strategic voting. Scand. J. Econ. 107, 175–195.
Burtraw, D., 1992. Strategic delegation in bargaining. Econom. Lett. 38, 181–185.
Burtraw, D., 1993. Bargaining with noisy delegation. RAND J. Econ. 24, 40–57.
Carraro, C., Siniscalco, D., 1993. Strategies for the international protection of the environment. J. Public Econ. 52, 309–328.
Christiansen, N., 2013. Strategic delegation in a legislative bargaining model with pork and public goods. J. Public Econ. 97, 217–229.
Crawford, V.P., Varian, H.R., 1979. Distortion of preferences and the Nash theory of bargaining. Econom. Lett. 3, 203–206.
Dur, R., Roelfsema, H., 2005. Why does centralisation fail to internalise policy externalities? Public Choice 122 (3), 395–416.
Finus, M., 2001. Game Theory and International Environmental Cooperation. Edward Elgar, Cheltenham.
Finus, M., Maus, S., 2008. Modesty may pay. J. Public Econ. Theory 10, 801–826.
Gersbach, H., Winkler, R., 2011. International emission permits markets with refunding. Eur. Econ. Rev. 55, 759–773.
Habla, W., Winkler, R., 2018. Strategic delegation and international permit markets: Why linking may fail. J. Environ. Econ. Manag. 92, 244–250.
Hagen, A., Altamirano-Cabrera, J.-C., Weikard, H.-P., 2020. National political pressure groups and the stability of international environmental agreements. Int. Environ. Agreem.: Polit. Law Econ. http://dx.doi.org/10.1007/s10784-020-09520-5.
Harstad, B., 2010. Strategic delegation and voting rules. J. Public Econ. 94, 102–113.
Harstad, B., 2020. Pledge-and-Review Bargaining: From Kyoto to Paris. Mimeo.
Hattori, K., 2010. Strategic voting for noncooperative environmental policies in open economies. Environ. Resour. Econ. 46, 459–474.
Helm, C., 2003. International emissions trading with endogenous allowance choices. J. Public Econ. 87, 2737–2747.
Hoel, M., 1992. International environment conventions: The case of uniform reductions of emissions. Environ. Resour. Econ. 2, 141–159.

Hoel, M., Schneider, K., 1997. Incentives to participate in an international environmental agreement. Environ. Resour. Econ. 9, 153–170.
Jones, S.R.G., 1989. Have your lawyer call my lawyer: Bilateral delegation in bargaining situations. J. Econ. Behav. Organ. 11, 159–174.
Karp, L., Simon, L., 2013. Participation games and international environmental agreements: A non-parametric model. J. Environ. Econ. Manag. 65, 326–344.
Kempf, H., Rossignol, S., 2013. National politics and international agreements. J. Public Econ. 100, 93–105.
Köke, S., Lange, A., 2017. Negotiating environmental agreements under ratification constraints. J. Environ. Econ. Manag. 83, 90–106.
Kopel, M., Pezzino, M., 2018. Strategic delegation in oligopoly. In: Corchón, L., Marini, M. (Eds.), Handbook of Game Theory and Industrial Organization. Edward Elgar, Chapter 10.
Loeper, A., 2017. Cross-border externalities and cooperation among representative democracies. Eur. Econ. Rev. 91, 180–208.
Marchiori, C., Dietz, S., Tavoni, A., 2017. Domestic politics and the formation of international environmental agreements. J. Environ. Econ. Manag. 81, 115–131.
Perino, G., 2010. How delegation improves commitment. Econom. Lett. 106, 137–139.
Persson, T., Tabellini, G., 1992. The politics of 1992: Fiscal policy and European integration. Rev. Econom. Stud. 59, 689–701.
Redoano, M., Scharf, K.A., 2004. The political economy of policy centralization: direct versus representative democracy. J. Public Econ. 88, 799–817.
Roelfsema, H., 2007. Strategic delegation of environmental policy making. J. Environ. Econ. Manag. 53, 270–275.
Schmalensee, R., 1998. Greenhouse policy architecture and institutions. In: Nordhaus, W.D. (Ed.), Economics and Policy Issues in Climate Change. In: Resources for the Future, Chapter 5.
Segendorff, B., 1998. Delegation and threat in bargaining. Games Econom. Behav. 23, 266–283.
Siqueira, K., 2003. International externalities, strategic interaction, and domestic politics. J. Environ. Econ. Manag. 45, 674–691.
Sobel, J., 1981. Distortion of utilities and the bargaining problem. Econometrica 49, 597–619.
Strøm, K., 2000. Delegation and accountability in parliamentary democracies. Eur. J. Polit. Res. 37, 261–290.
Wagner, U.J., 2001. The design of stable international environmental agreements: economic theory and political economy. J. Econ. Surv. 15, 377–411.
de Zeeuw, A., 2015. International environmental agreements. Annu. Rev. Resour. Econ. 7, 151–168.