b
UNIVERSITÄT
BERN

Political Economy Factors in International Environmental Cooperation

Author
Sarah SPYCHER

Supervisor
Prof. Dr. Ralph WINKLER

Inaugural Dissertation in Fulfillment of the Requirements for the Degree of

DOCTOR RERUM OECONOMICARUM

at the

Department of Economics
Faculty of Business, Economics and Social Sciences

February 3, 2023

Die Fakultät hat diese Arbeit am 30. März 2023 auf Antrag der beiden Gutachter Prof. Dr. Ralph Winkler und Prof. Dr. Andreas Lange als Dissertation angenommen, ohne damit zu den darin ausgesprochenen Auffassungen Stellung nehmen zu wollen.

Acknowledgements

First and foremost I would like to thank my supervisor Ralph Winkler, for all the support and confidence in me ever since we started working together on my Master's thesis. I have immensely benefited from your expertise, ideas and enthusiasm over the course of the years and I could always rely on your help and your presence, be it online through a global pandemic and months spent abroad, or just down the corridor. Thank you Ralph for supervising my thesis and for being my mentor, I have always thoroughly enjoyed working with you. I would also like to thank Andreas Lange for generously accepting to evaluate my thesis as external reviewer.

I am very grateful for my wonderful office mates in A227: Philipp, Anna, Leyla, Moritz and Severin. I truly enjoyed the fun and lighthearted atmosphere chatting away and playing darts, while always allowing for productive and motivating discussions. The same goes for Yell and Carla, almost honorary members of our office and also very dear to my heart. A big thank you to all my PhD companions, current as well as previous, in particular Sarina Steinmann, Gala Sipos, Nadia Ceschi, Tamara Bischof and Jonas Meier. All of you inspired me in your own special ways and became friends. I would also like to express my gratitude to all the current and previous members of the Department's administrative office – it was lovely to work with all of you and I always enjoyed our little chats.

I have been fortunate enough to be part of the Micro group within our department, which over the years has grown into such a helpful resource. This group has given me a sense of community and team spirit in a sometimes lonely endeavour as a PhD student. Special thanks to Marc Möller and Jean-Michel Benkert, who very generously supported me in writing grant and job applications in the last two years. I would also like to extend my deepest gratitude to Catherine Roux, who served as my mentor early on in my PhD and has been an avid supporter ever since – thank you for agreeing to participate in the mentoring programme, and for being a role model.

I was fortunate enough to spend 6 months at the Department of Geography and Environment at the London School of Economics during my PhD, thanks to funding through the SNSF Doc.Mobility scheme and the hospitality of my host professor Simon Dietz. This research stay was very inspirational on many levels, but mainly because of the connections I made. A big thank you to all my PhD colleagues there and a special shout-out to Giulia Romani, Sanchayan Banerjee and Manuel Linsenmeier. Seeing you so fearlessly pursuing ambitious goals, all the while being so much fun to be around, truly left a mark on me.

Of course I am deeply indebted to my family and friends for all their encouragement and patience. First, my parents who always had my back throughout the ups and downs of the past years, and a special mention to my Grossvati, who has followed my academic journey like few others. Second, to all my close friends for their unwavering support and for taking my mind off things when I needed it the most, in particular Tabea, Raphi, Chloé, Gina, Vanessa and Steffi. Finally, my deepest appreciation goes to my partner Tobi, for being my safe haven and my biggest fan. Your love has been the greatest support.

Contents

List of Figures

List of Tables

Introduction

Scientifically, it is undisputed that decisive and immediate action is needed to achieve the goal of limiting the impacts of climate change (IPCC 2022). However, little progress on climate change mitigation can be observed: current mitigation efforts by countries, as pledged in the Paris Agreement, are not ambitious enough to meet the recognised policy goal of containing the increase in average surface temperature below 2°C compared to pre-industrialised levels. In addition, in almost all countries, current greenhouse gas emissions are on a higher path than pledged (UNEP 2022).

From an economic point of view, the lack of success in international cooperation on environmental policy is not surprising, since mitigation of climate change is impeded by the public goods property of greenhouse gas emission reductions. This makes climate change a transboundary issue that has to be jointly tackled by the global community, which requires international coordination and cooperation, a task hindered by the fact that no supranational entity can enforce a fair and efficient outcome. Unsurprisingly, achieving a consensus that is significant and impactful has proven to be extremely difficult over the past decades and we are observing a constant underprovision of emission reductions globally due to freeriding incentives. Both the Kyoto Protocol and its successor, the Paris Agreement indisputably demonstrate the difficulties of achieving ambitious environmental agreements as well as the reluctance of participating countries to comply with emission targets agreed upon.

The discipline of economics has contributed to the solution of this problem by designing treaty mechanisms and policy tools for the efficient implementation of ambitious environmental goals. However, the primary challenge does not mainly lie in the scientific (geophysical) consensus nor in the lack of suitable policy options, but more so in the problem of global burden sharing as well as the inertia of political systems across the globe (Matthews and Wynes 2022).

Environmental policy is very strongly linked to the political system in which policymaking takes place. The nature of the political processes at hand can therefore directly interfere with or support the success of such policies. As a result of the hierarchical structure of international climate policy, political economy considerations concerning the incentives and disincentives for participation and the design in international environmental agreements could be among the key explanatory factors for the current

underprovision of effective international cooperation (Battaglini and Harstad 2020). Both the international and the domestic level of environmental policy, if limited by political economy factors, bear the risk of putting collective efforts in the mitigation of climate change to a halt.

Even though these challenges are highly topical, present in current policy debates and a lot is at stake for the global community, the interaction between political economy and international environmental cooperation has received comparably little attention in the economic literature so far. Gaining a deeper understanding of factors that are hindering more successful cooperation is crucial and can be tackled from different perspectives. Theoretical models allow us to investigate the principal incentive structures underlying agents' decisionmaking processes. Another approach lies in experimental studies: since their setup is conceptually close to the negotiation of international environmental agreements, they can offer valuable insights into how agents behave in a variety of circumstances. In my thesis I employ both methods, addressing the overarching topic of how political economy factors impact environmental policy and how they are hindering more decisive actions in mitigating climate change.

This thesis comprises of two theoretical chapters addressing the hierarchical structure of environmental policy and the understanding of how this affects international environmental cooperation. The third chapter is an experimental analysis, offering a perspective on how heterogeneous stakeholders interact in the provision of a common public good, such as international cooperation on emission reductions. The common thread of the three chapters lies in the question of how the structure and decisionmaking process in international environmental policy facilitates or hinders the provision of public goods.

In the first chapter of my thesis, co-authored with Ralph Winkler and published in the *European Economic Review*, we revisit a prominent approach for the analysis of so-called self-enforcing international environmental agreements (IEAs), designed such that it is in each country's best interest to comply with the treaty and thus, given the absence of a supranational authority, no enforcing entity is necessary. The formation and stabilization of IEAs is then analysed with non-cooperative game theoretical models. Most standard model frameworks regard each country as a homogeneous entity. We, however, argue that modern democracies typically feature a chain of delegation from voters to those who govern and decide on participation in IEAs. To account for this structure, we introduce a principal-agent framework. We find that the institutional setup assumed, that is the timing of the game and thus the way in which decision power is split between principal and agent, crucially determines the circumstances that make successful IEAs more or less difficult to achieve. We show that institutions play a decisive role in the design of international policy: delegation may act as a commitment device to provide an option for more credible signalling of intentions to cooperate.

This paper adds a number of novel insights to the existing literature. On the one hand, we reassess the omnipresent "narrow-but-deep" versus "broad-but-shallow" trade-off (e.g. Barrett 2002; Finus and Maus 2008), that is agreements are either ambitious in their effective public good provision but only consist of a small number of participating parties, or supported by many involved parties, which comes at a sharp reduction in the effective provision of the public good of each signatory. We show

that this trade-off does not exist in our strategic delegation coalition formation framework and that the setup allows for "broad-and-deep" agreements. On the other hand, we identify two different motives to strategically delegate, the first of which stems from the strategic substitutability of emission choices and is well understood in the literature. The second delegation motive, however, is unique to the modest coalition formation model with strategic delegation: by delegating to a "greener" agent, principals can increase the "effective" environmental damage internalisation and therefore counteract any deviation from an ambitious agreement. Lastly, we refine the framework of modest coalition formation games, which raises serious concerns with respect to compliance, by providing an institutional microfoundation for the emission choice stage based on the international permit market with refunding mechanism introduced by Gersbach and Winkler (2011).

In my second chapter, I investigate the role that domestic elections play for IEAs and to what extent they may be an explanatory factor for the modest success of international cooperation on climate change mitigation. In doing so, I pay particular attention to the role of political polarisation within a country. Agents involved in international negotiations are often subject to domestic electoral concerns and therefore, policy decisions might affect their chances of re-election in national elections. Also, international treaties usually last beyond a governments' incumbency. This leads to a temporal disparity in the sense that environmental treaties are usually devised to last over a long period of time, while election cycles are comparably short. This may also imply that the negotiation and the ratification decision might be made by two different governments. With these considerations in mind, I formulate a 4-stage game modelling a bilateral environmental agreement and find that incumbent governments often negotiate treaties different from what they would choose in the absence of electoral competition.

This distortion can occur in a number of different ways: incumbents may negotiate (i) a *consensus treaty*, anticipating that they might be replaced in an upcoming election and devising the agreement such that their successor would still ratify, (ii) a *differentiation treaty* which exaggerates their preferences such that the agreement is too extreme for the challenger to sign and therefore the two parties offer differing environmental policies to the median voter, or lastly (iii) an *insurance treaty* negotiating an agreement that neutralises the potential successor with a predetermined, tamed policy path. From the perspective of a median voter, all of the abovementioned treaty types can lead to a distorted outcome which negatively affects their welfare. Political polarisation, an aspect so far underexplored in the literature, tends to accentuate this negative impact. Overall, this paper thus highlights the importance of considering domestic electoral competition, including political polarisation, in the analyisis of why we observe a constant underprovision of public goods such as emission reductions.

In the third chapter, I conduct an experimental study of a *threshold public goods game* in which players have varying preferences for the public good. In general, public goods games offer a stylised insight into a collective action problem, with the goal of shedding light on the relative importance of factors influencing the success of cooperation. While my experimental framework is not specific to any application, it can be interpreted in an environmental context: agents involved in the negotiation of IEAs

offer contributions to a common public good, as for example averting collective damage from climate change, where a certain level of investments is necessary to make a treaty stringent and useful. The players involved often are heterogeneous with respect to their stakes, which can be interpreted as them valuing successful provision of the public good differently depending on, for example, their exposure or vulnerability to climate damages. As a consequence, given the fact that the setup is conceptually close to the negotiation of international treaties, results from this experiment might prove useful in informing, for example, future UNFCCC negotiations or bilateral international environmental cooperation.

In my specific experimental design, I pay particular focus on the effect of different preferences for the public good on (i) the frequency of provision, (ii) the equilibrium selection, and (iii) how players split the contribution costs in the presence of different degrees of asymmetry. I find that groups with symmetric players have a significantly higher success rate in reaching the threshold than asymmetric players. Surprisingly however, the specific degree of asymmetry has almost no effect on provision success. Furthermore, when considering burden-sharing, while players notice and act on asymmetric benefits, the degree of asymmetry is less salient than hypothesised. This leads to outcomes in which "poor" players contribute too much while "rich" players contribute too little, diverging from both our benchmark allocations of efficiency or fairness. Due to the existence of large differences with regards to wealth and exposure to climate change damages across the world, equity considerations play an important role in successful international environmental policy. Re-contextualising my experimental findings would thus imply that richer countries might underestimate the degree to which their contributions should exceed that of poorer nations when aiming for equitable outcomes.

The overarching aim of this thesis is to contribute to a more profound comprehension of potential obstacles and prospects of future international environmental cooperation by accounting for political economy frictions not previously or sufficiently researched. Gaining a deeper understanding of the effects of domestic electoral competition combined with political polarisation and the associated incentives for governments in power, delegation mechanisms within domestic political systems as well as burden-sharing among heterogeneous stakeholders will prove necessary in order to provide instructive policy guidance for the design of more effective environmental policy in the future.

References

Barrett, S. (2002). Consensus treaties. *Journal of Institutional and Theoretical Economics*, 529–547.

Battaglini, M. and B. Harstad (2020). The political economy of weak treaties. *Journal of Political Economy 128*(2), 544–590.

Finus, M. and S. Maus (2008). Modesty may pay. *Journal of Public Economic Theory 10*, 801–26.

Gersbach, H. and R. Winkler (2011). International emission permit markets with refunding. *European Economic Review 55*(6), 759–773.

IPCC (2022). Climate Change 2022: Mitigation of Climate Change. Contribution of Working Group iii to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. *[P.R. Shukla, J. Skea, R. Slade, A. Al Khourdajie, R. van Diemen, D. McCollum, M. Pathak, S. Some, P. Vyas, R. Fradera, M. Belkacemi, A. Hasija, G. Lisboa, S. Luz, J. Malley, (eds.)]*.

Matthews, H. D. and S. Wynes (2022). Current global efforts are insufficient to limit warming to 1.5°C. *Science 376*(6600), 1404–1409.

UNEP (2022). Emissions gap report 2022: The closing window – climate crisis calls for rapid transformation of societies. *https://www.unep.org/emissions-gap-report-2022*.

Chapter 1

Strategic Delegation in the Formation of Modest International Environmental Agreements

Published as:

Spycher, S. and R. Winkler (2022). Strategic Delegation in the Formation of Modest International Environmental Agreements. *European Economic Review 141*, 103963.
https://doi.org/10.1016/j.euroecorev.2021.103963

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.

1.1 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 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

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

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 knowl-

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

edge, 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.

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

[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).

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

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

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).

1.2 The Model

We consider a set $I = \{1, \ldots, n\}$ 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.

1.3 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 homogenous 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.

1.3.1 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 1.9, 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]

[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.

1.3.2 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 1.9.

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.

1.4 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 $\Theta = (\theta_1, \ldots, \theta_n)$ detailing the preference parameters of the selected agents in all countries. Agents in non-member countries $i \notin S$ maximize V_i:

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

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 $\Theta = (\theta_1, \ldots, \theta_n)$ 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 $\hat{e}(S,\Theta) = \big(\hat{e}_1(S,\Theta),\ldots,\hat{e}_n(S,\Theta)\big)$ 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.

1.5 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.

1.5.1 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 equation (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 Equation (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 Equations (13), which determine

[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.

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 Equations (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 Equation (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.

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 Figure 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

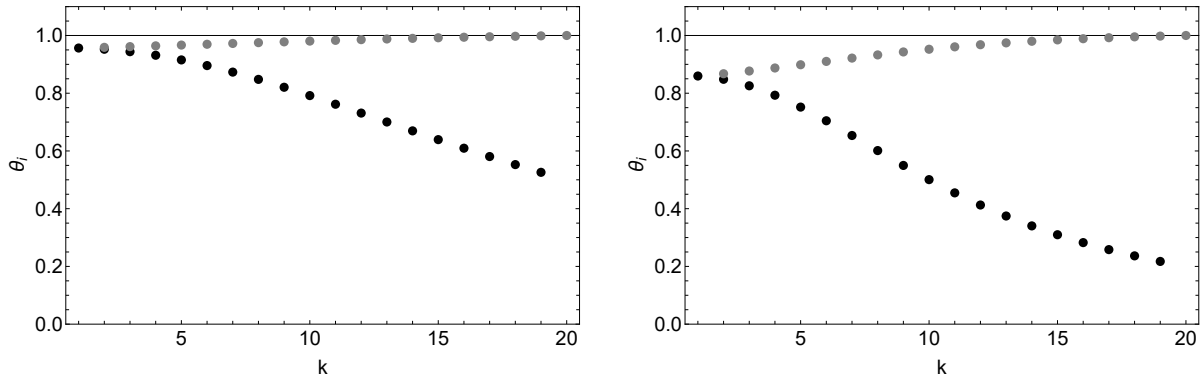


Figure 1: Illustration of θ_i^{NS} (black dots) and θ_i^S (gray dots) as a function of the stable coalition size k for $n = 20$ and $\phi = 0.025$ (left) and $\phi = 0.01$ (right). For $k = 1$, θ_i^{NS} and θ_i^S coincide. For increasing k, θ_i^S increases while θ_i^{NS} decreases. For the grand coalition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

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.

1.6 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 1.5.1, 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 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.

1.7 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 1.6, i.e., $\epsilon = \beta = 1$ and $\phi = \delta$. In the strong delegation set-up,

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

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:

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

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

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

1.7.1 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 1.6, 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 $\hat{k} \in \{k^{min},\ldots,k^{max}\}$. 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 $\{k^{min}(n,\gamma),\ldots,k^{max}(\gamma)\}$ 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.

[9] We shall confirm this in Section 1.7.2.

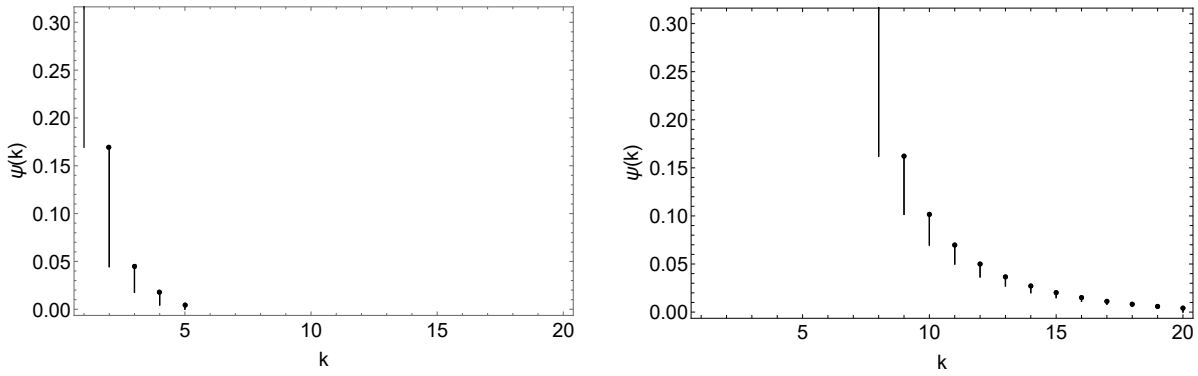


Figure 2: Illustration of function $\psi(k)$ for $n = 20$ and $\gamma = 0.6$ (left) and $\gamma = 0.15$ (right). The dots indicate the values for $\psi(k)$ obtained by inserting $k \in \{k^{min}, \ldots, k^{max}\}$. The lines indicate the stable coalition sizes $\hat{k}$ for values of ψ between $\psi(k-1)$ 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 $\psi = \phi\theta$.

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., $\psi(k)_{k \to \underline{k}, k > \underline{k}} = +\infty$. Figure 2 shows $\psi(k)$ for two different values of γ.

Note that the graphs in Figure 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.

1.7.2 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

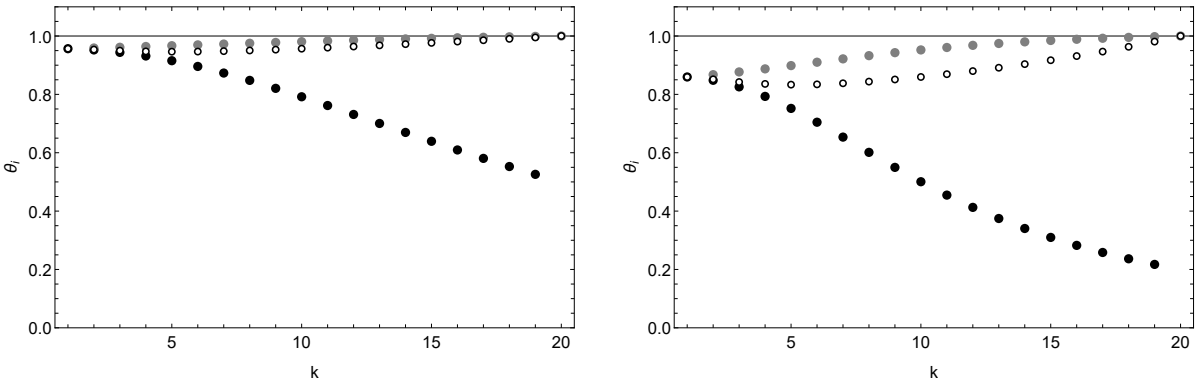


Figure 3: Illustration of $\hat{\theta}$ (black circles) in the strong delegation game compared to θ_i^{NS} (black dots) and θ_i^S (gray dots) in the weak delegation game as a function of the stable coalition size k for $n = 20$, $\gamma = 1$, and $\phi = 0.025$ (left) and $\phi = 0.01$ (right). For $k = 1$, $\hat{\theta}$, 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.

simultaneously delegate to agents such as to maximize their *expected welfare*:

$$\begin{aligned} \max_{\theta_i} \quad & \frac{k(\Theta)}{n}\Big(B\big(e_i^S(\Theta,k(\Theta))\big) - D\big(E(\Theta,k(\Theta))\big)\Big) \\ & + \frac{n-k(\Theta)}{n}\Big(B\big(e_i^{NS}(\Theta,k(\Theta))\big) - D\big(E(\Theta,k(\Theta))\big)\Big)\ . \end{aligned} \tag{21}$$

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 1.5.1, 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.

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 equilib-

rium 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 $i \in I$:

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

Equation (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 equation (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 Figure 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

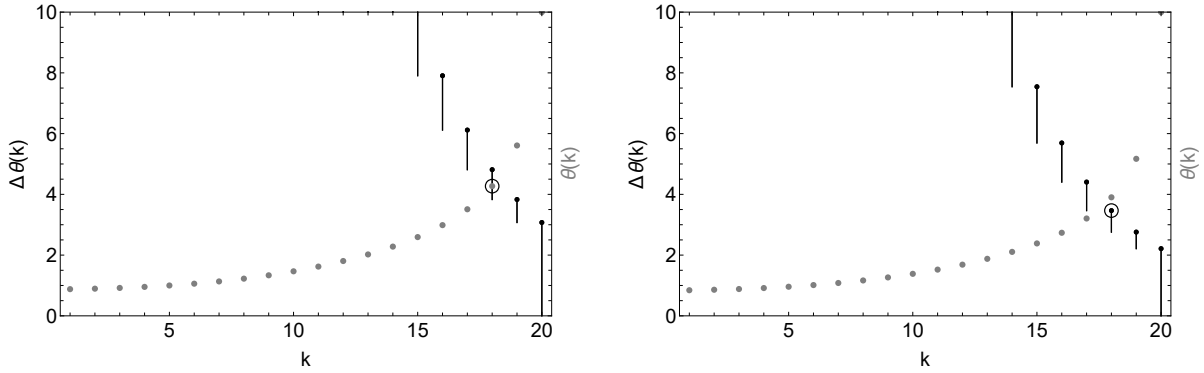


Figure 4: Illustration of $\Delta\theta(k)$ (black) and $\hat{\theta}(k)$ (gray), which determine the subgame perfect equilibrium (SPE) of the strong delegation game, for $n = 20$ and $\gamma = 0.1$. An "interior"-SPE occurs if $\Delta\theta(k)$ and $\hat{\theta}(k)$ "intersect" (left). If they do not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

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 Figure 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]

[10] Figure 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.

1.8 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 1.5.1, 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 Figure 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?

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.

[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.

1.9 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,

[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.

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 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

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

– 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

[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.

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.

Acknowledgements

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" 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.1.1 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 $\Theta = (\theta_1, \ldots, \theta_n)$ 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 $\Omega = (\omega_1, \ldots, \omega_k)$ 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:

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

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:

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

which leads to the equilibrium permit price:

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

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

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

Relationship between μ and γ

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 equation (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.1.2 Proof of Proposition 1

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

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

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

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 equations (A.13) yields the unique Nash equilibrium $\hat{e}(S,\Theta)$. □

A.1.3 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.1.4 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 equations (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 Figure 5. □

A.1.5 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.1.6 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 equations (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

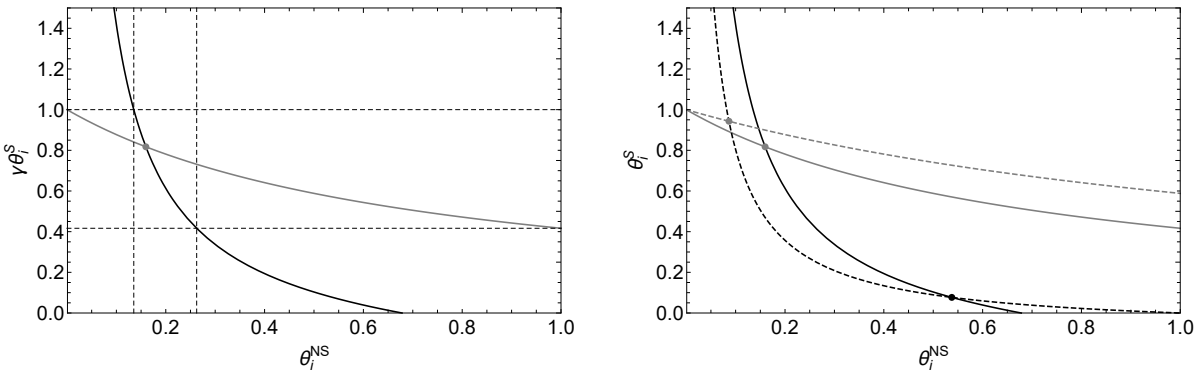


Figure 5: Illustration of the proofs of Proposition 3 (left) and Proposition 5 (right). The left plot shows the intersection of the reaction functions (R_1 in black and R_2 in gray), which exists and is unique. The horizontal lines show bounds for the feasible range for $\gamma\hat{\theta}_i^S$ in equilibrium, 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.

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:

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

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 Figure 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 equation (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.1.7 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.1.8 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 1.6, 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:

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

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:

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

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 $\hat{U}_i^{NS}\left(k-1,\theta_i^S(k-1),\theta_i^S(k-1)\right) < \hat{U}_i^{NS}\left(k-1,\theta_i^{NS}(k-1),\theta_i^S(k-1)\right)$.

(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:

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

[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 $\theta_i^S = 1/[1+\phi(n-k)\theta_i^{NS}]$.

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

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

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:

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

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.1.9 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):

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

$Z(k,\theta) \lesseqgtr 0$ 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

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

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

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

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

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:

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

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:

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

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 equation (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 equation (A.54). Thus, we obtain:

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

□

A.1.10 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 contradition. 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 quilibrium:

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

This is equivalent to:

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

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 equations (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:

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

For $k = n$ this reduces to

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

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:

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

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:

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

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:

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

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

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

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:

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

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.1.11 Proof of Proposition 10

Recall from equations (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.1.12 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.1.13 Robustness Check: Heterogenous 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 heterogenous 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 $i \in \mathcal{S}$ 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 1.9, 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 Equation (A.87b), all distributions of θ_i among member countries that satisfy Equation (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 Equation (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 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 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 equations (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.1.14 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.

[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.

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

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 $W^\star$.

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., S. Barrett, and R. N. Stavins (2003). Thirteen plus one: a comparison of global climate policy architectures. *Climate policy 3*, 373–397.

Barrett, S. (2002). Consensus treaties. *Journal of Institutional and Theoretical Economics 158*, 529–547.

Barrett, S. (2003). *Environment and Statecraft: The Strategy of Environmental Treaty-making*. Oxfrod University Press.

Battaglini, M. and B. Harstad (2020). The political economy of weak treaties. *Journal of Political Economy 128*, 544–590.

Besley, T. and S. Coate (2003). Centralized versus decentralized provision of local public goods: a political economy approach. *Journal of Public Economics 87*(12), 2611–2637.

Buchholz, W., A. Haupt, and W. Peters (2005). International environmental agreements and strategic voting. *Scandinavian Journal of Economics 107*, 175–195.

Burtraw, D. (1992). Strategic delegation in bargaining. *Economics Letters 38*, 181–185.

Burtraw, D. (1993). Bargaining with noisy delegation. *The RAND Journal of Economics 24*, 40–57.

Carraro, C. and D. Siniscalco (1993). Strategies for the international protection of the environment. *Journal of Public Economics 52*, 309–28.

Christiansen, N. (2013). Strategic delegation in a legislative bargaining model with pork and public goods. *Journal of Public Economics 97*, 217–229.

Crawford, V. P. and H. R. Varian (1979). Distortion of preferences and the Nash theory of bargaining. *Economics Letters 3*, 203–206.

de Zeeuw, A. (2015). International environmental agreements. *Annual Reviews of Resource Economics 7*, 151–168.

Dur, R. and H. Roelfsema (2005). Why does centralisation fail to internalise policy externalities? *Public Choice 122*(3), 395–416.

Finus, M. (2001). *Game Theory and International Environmental Cooperation*. Cheltenham: Edward Elgar.

Finus, M. and S. Maus (2008). Modesty may pay. *Journal of Public Economic Theory 10*, 801–26.

Gersbach, H. and R. Winkler (2011). International emission permits markets with refunding. *European Economic Review 55*, 759–73.

Habla, W. and R. Winkler (2018). Strategic delegation and international permit markets: Why linking may fail. *Journal of Environmental Economics and Management 92*, 244–250.

Hagen, A., J.-C. Altamirano-Cabrera, and H.-P. Weikard (2020). National political pressure groups and the stability of international environmental agreements. *International Environmental Agreements: Politics, Law and Economics*. https://doi.org/10.1007/s10784-020-09520-5.

Harstad, B. (2010). Strategic delegation and voting rules. *Journal of Public Economics 94*, 102–113.

Harstad, B. (2020, June). Pledge-and-review bargaining: From Kyoto to Paris. Mimeo.

Hattori, K. (2010). Strategic voting for noncooperative environmental policies in open economies. *Environmental and Resource Economics 46*, 459–474.

Helm, C. (2003). International emissions trading with endogenous allowance choices. *Journal of Public Economics 87*, 2737–2747.

Hoel, M. (1992). International environment conventions: The case of uniform reductions of emissions. *Environmental and Resource Economics 2*, 141–59.

Hoel, M. and K. Schneider (1997). Incentives to participate in an international environmental agreement. *Environmental and Resource Economics 9*, 153–70.

Jones, S. R. G. (1989). Have your lawyer call my lawyer: Bilateral delegation in bargaining situations. *Journal of Economic Behavior & Organization 11*, 159–174.

Karp, L. and L. Simon (2013). Participation games and international environmental agreements: A non-parametric model. *Journal of Environmental Economics and Management 65*, 326–44.

Kempf, H. and S. Rossignol (2013). National politics and international agreements. *Journal of Public Economics 100*, 93–105.

Köke, S. and A. Lange (2017). Negotiating environmental agreements under ratification constraints. *Journal of Environmental Economics and Management 83*, 90–106.

Kopel, M. and M. Pezzino (2018). Strategic delegation in oligopoly. In L. Corchón and M. Marini (Eds.), *Handbook of Game Theory and Industrial Organization*, Chapter 10. Edward Elgar.

Loeper, A. (2017). Cross-border externalities and cooperation among representative democracies. *European Economic Review 91*, 180–208.

Marchiori, C., S. Dietz, and A. Tavoni (2017). Domestic politics and the formation of international environmental agreements. *Journal of Environmental Economics and Management 81*, 115–131.

Perino, G. (2010). How delegation improves commitment. *Economics Letters 106*, 137–139.

Persson, T. and G. Tabellini (1992). The politics of 1992: Fiscal policy and European integration. *The Review of Economic Studies 59*, 689–701.

Redoano, M. and K. A. Scharf (2004). The political economy of policy centralization: direct versus representative democracy. *Journal of Public Economics 88*(3), 799–817.

Roelfsema, H. (2007). Strategic delegation of environmental policy making. *Journal of Environmental Economics and Management 53*, 270–275.

Schmalensee, R. (1998). Greenhouse policy architecture and institutions. In W. D. Nordhaus (Ed.), *Economics and Policy Issues in Climate Change*, Chapter 5. Resources for the Future.

Segendorff, B. (1998). Delegation and threat in bargaining. *Games and Economic Behavior 23*, 266–283.

Siqueira, K. (2003). International externalities, strategic interaction, and domestic politics. *Journal of Environmental Economics and Management 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. *European Journal of Political Research 37*, 261–290.

Wagner, U. J. (2001). The design of stable international environmental agreements: economic theory and political economy. *Journal of Economic Surveys 15*, 377–411.

Chapter 2

Elections, Political Polarisation and Environmental Agreements

Abstract: This paper investigates the role that domestic elections play for IEAs and to what extent they might be an explanatory factor for the modest success of recent international cooperation on climate change mitigation. Agents involved in international negotiations are often subject to domestic electoral concerns and therefore, policy decisions might affect their chances of reelection in upcoming elections. Also, international treaties usually last beyond a governments' incumbency, which implies that the negotiation and the ratification decision might be made by two different entities. I formulate a 4-stage game modelling a bilateral environmental agreement in order to analyse the arising strategic incentives depending on the level of political polarisation. I find that incumbent governments often choose suboptimal treaties (compared to if there was no election) in order to boost their chances of reelection. This can happen in three different ways: the incumbent may negotiate (i) a "consensus treaty", anticipating that they might be replaced in an upcoming election and devising the agreement such that their successor would still ratify, (ii) a "differentiation treaty" such that the two parties offer differing environmental policies to the voters in the election, or lastly (iii) an "insurance treaty", negotiating an agreement devised to serve as a safeguard against the challenger in case the election is lost. Increased polarisation generally leads to more distorted treaties and worse outcomes from the perspective of the median voter.

2.1 Introduction

Anthropogenic climate change is widely recognised as one of the major global environmental issues of our times. In the past few decades, the international community has addressed the subject by negotiating many international environmental agreements (IEAs), most recently resulting in the *Paris Agreement* in 2015. However, little progress on climate change mitigation can be observed: the current pledges as agreed upon in the Paris Agreement are not ambitious enough to meet the recognised policy goal of keeping the increase in average surface temperature below 2°C below pre-industrialised levels. In addition, in almost all countries, current greenhouse gas (GHG) emissions are on a higher path than pledged (UNEP 2022).

Transnational cooperation on climate change mitigation poses a fundamental challenge to the international community: both the Paris Agreement and its predecessor the *Kyoto Protocol* indisputably demonstrate the difficulties of achieving ambitious environmental agreements as well as the reluctance of participating countries to comply with emission targets agreed upon. This lack of success is not surprising from an economic point of view: on the one hand, mitigation of anthropogenic climate change is impeded by the public goods property of GHG emission reductions. Each country's efforts to reduce emissions benefits all countries in a non-exclusive and non-rival manner, while costs are borne domestically. At the same time, no supranational authority exists that might enforce an efficient outcome. We therefore observe a global underprovision of emission reductions.

This paper investigates the role that domestic elections play for IEAs and to what extent they might be an explanatory factor for the modest success of current international cooperation on climate change mitigation. Agents involved in international negotiations are often subject to domestic electoral concerns and therefore, policy decisions might affect their chances of reelection in upcoming elections. Secondly, international treaties usually last beyond a government's incumbency. This, on the one hand, leads to a temporal disparity in the sense that environmental treaties are generally devised to last over a long period of time, while election cycles are comparably short. On the other hand, this implies that the negotiation and the ratification decision might be made by two different entities.

A good example of such deliberations is the behaviour of the US in the negotiation process of the Kyoto protocol. Al Gore, acting as the vice president in the Clinton administration took part in negotiating a – from the point of view of the US – ambitious target. However, the administration was well aware of the fact that the ratification of such a treaty would be rejected by the Senate. One could argue that Al Gore, planning to run for president in the upcoming election strategically positioned himself on the topic of environmental policy in order to positively influence his electoral prospects. This example also serves as an argument in favour of separating the negotiation and ratification decision in the model, since it might be made by two different entities.

With these considerations in mind, I formulate a 4-stage game modelling a bilateral environmental agreement in order to analyse the arising strategic incentives. Two countries set up a bilateral agreement on emission reductions, where I focus on political competition within country 1. In a first stage,

the incumbent in country 1 negotiates a treaty, accounting for the fact that its contents affect their re-election chances in an upcoming domestic election. After the election, the outcome of which depends on the median voter's welfare and a stochastic shock, whoever is in government chooses to ratify the negotiated treaty or not. Emission choices follow as a result of the ratification decision.

Results indicate that incumbent governments might indeed choose a "suboptimal" treaty (compared to if there was no election) in order to boost their chances of reelection, driven by a number of political economy factors and varying degrees of political polarisation. This can happen in a number of different ways: the incumbent may negotiate a "consensus treaty", anticipating that they might be replaced in an upcoming election and devising the agreement such that their successor would still ratify, which is usually to the benefit of the median voter. Also, a "differentiation treaty" can emerge, meaning that the two parties offer differing environmental policies to the voters in the election, which especially in case of high polarisation, are a bad option for the median voter. Lastly, an "insurance treaty" is possible, where an agreement devised to serve as a backstop against the challenger in case the election is lost, a type of treaty that is very much to the detriment of the median voter. We also discuss the few limited cases, in which the first-best outcome, that is, the outcome in the absence of elections, would materialise. Finally and as a novel insight I demonstrate the importance of considering domestic political polarisation in the discussion of international cooperation. This goes beyond the topic of environmental policy: I can illustrate how in a two party system, the provision of a shared public good in general becomes more difficult with polarised parties. Given the fact that the US, a prominent example of a two party democracy, is one of the major global players when it comes to international cooperation, this connection will become increasingly relevant.

2.2 Related Literature

The question of how elections affect policy choices, and vice versa, has been widely discussed in contexts outside of international (environmental) cooperation before. Persson and Tabellini (1992) show that political processes such as elections may distort tax rate choices compared to what a social planner would do, while Besley and Coate (1998) highlight how fiscal policy investments can be used to influence future elections. Robinson and Torvik (2005) show how inefficient investments in local infrastructures might stem from attempts to influence elections. However, in Persson and Tabellini (1994), contrary to a majority of the public choice literature, the authors find that political incentives may also improve the equilibrium outcome through more credible commitment. Addressing how incumbent governments can influence policies of their successors, Alesina and Tabellini (1990) discuss the role of public debt as a means of limiting expenditures. My paper contributes to this literature by adding insights into the nexus of economic policy and political competition in the context of cross-border public goods provision, specifically in an environmental context.

In the environmental economics literature, the theoretical analysis of *self-enforcing* IEAs is usually based on models of non-cooperative game theory. In particular, coalition formation games have been

used to model IEAs since the early 1990s. The standard coalition formation game consists of two stages: in the membership stage, countries decide on whether to become a signatory of the agreement. In the subsequent emission choice stage, non-signatories play their respective non-cooperative Nash equilibrium, while signatory countries choose their domestic emission levels such as to internalize externalities from emissions within the coalition (Wagner 2001). This structure has been used as the basis for many models, which generally draw pessimistic conclusions about successful international cooperation: whenever gains from cooperation would be large, resulting coalition sizes are small and thus coalitions achieve little (e.g., Carraro and Siniscalco 1993, Hoel 1992, Barrett 1994). However, while the standard literature predicts free-riding and small coalitions, the existence of a multitude of large IEAs has been seen as a puzzle. This observation is sometimes referred to as the *paradox of international environmental agreements* (Kolstad and Toman 2005).

One way of resolving this paradox is by considering so-called *modest* IEAs, as done in a model presented by Finus and Maus (2008). Modesty in this context describes the depth or ambitiousness of an agreement and refers to the fact that, potentially, only a fraction of emission externalities is internalized within the coalition. The authors show that more modest agreements lead to higher membership sizes. Also, even though each coalition member chooses higher domestic emission levels as compared to more ambitious agreements with equal coalition size, this negative effect is outweighed by the larger number of members. Further papers that explain larger but more "shallow" coalitions include for example Barrett (2002), Aldy et al. (2003), Harstad (2022).

All of the aforementioned environmental literature, including wide-ranging extensions, have in common that countries are modelled to be homogeneous entities, represented by a single, benevolent decision maker. With this view, however, potential interactions between domestic and international environmental policy are neglected (Finus 2008). A novel and growing strand of economic literature thus focusses on enriching the structure of players involved in international cooperation, thereby accounting for hierarchical structures. Based on the political science literature, in which the relationship between domestic and international policy is described in the context of a two-level game (Putnam 1988), the distinction between different governmental bodies within countries is emphasised. It is argued that a more realistic analysis of IEAs calls for the inclusion of specific political economy factors such as interest groups, electoral concerns or domestic political structures. For example, Marchiori et al. (2017) as well as Hagen et al. (2021) discuss the effect of legislative lobbying on the formation of IEAs. In the former model in particular, the authors explain why governments may want to use IEAs in order to improve their bargaining position with respect to powerful business lobbies. Furthermore, Spycher and Winkler (2022) show that by accounting for the hierarchical structure of international climate policy via the introduction of a strategic delegation stage, "broad-and-deep" agreements can be stabilised.

So far, there are only few papers that combine the analysis of international cooperation on environmental policy with political competition on a national level, four of which are particularly relevant to our paper. First, in a strategic voting model with uncertain median voter preferences, Köke and

Lange (2017) analyse the impact of ratification constraints on the outcome of IEAs. In particular, they consider countries consisting of multiple players, that is they differentiate between "representatives" who negotiate the agreement and "pivotal agents", who decide on ratification. The authors formulate a coalition formation game, where representatives decide on whether or not to participate in the negotiation process. In this process, the minimum number of ratifiers as well as the corresponding commitment level is specified before the pivotal agents in each country choose to ratify the treaty or not. The agreement only comes into force if the minimum participation constraint is met. The authors identify the political economy dynamics within countries as a driving force for the size and scope of climate agreements and thus offer a public choice motivation to the results of Finus and Maus (2008). By doing so, the importance of closely analysing relevant political economy aspects influencing the position of countries within international negotiations is highlighted.

Second, the model of Battaglini and Harstad (2020) addresses electoral considerations of governments in multilateral climate negotiations. In a simple two-country model, in which one country exerts an externality on the other, a treaty specifying abatement commitments and sanctions in case of non-compliance is negotiated. By choosing this simplified bilateral model framework over a standard coalition formation game, the authors aim at shifting the focus away from participation concerns and more towards treaty design. They show that incumbent governments strategically choose to sign "weak treaties", that is treaties with too low sanctions to guarantee compliance, in order to influence future elections in their favour. For example, a government with low environmental preferences signs a treaty involving sanctions low enough for the median voter to favour non-compliance and thus reelection of the incumbent. This incentive is particularly strong when benefits from staying in office are large.

Coming more from a political science angle, Buisseret and Bernhardt (2018) discuss the effect on the terms of international agreements when there is the prospect of electoral replacement. In a model consisting of two countries, they allow for renegotiation after the election has taken place by the challenging government. They analyse how the terms of an initial agreement affect the reelection chance for the incumbent as well as the bargaining attitude of a potential successor. Their main finding is that whether an agreement is ratified hinges on how friendly (towards the agreement) the incumbent is and on the timing of the election: if the incumbent is hostile, that is has lower environmental valuation, then the agreement is signed only if the election is relatively far away. In contrast, if the incumbent is friendly, then the treaty is ratified only if the election is sufficiently close.

Lastly, in a similar vein and most recently, Melnick and Smith (2022) discuss the interplay between elections and political dynamics in bilateral agreement bargaining. In a model setup resembling mine, yet without externalities of any sorts and abstracting from the impacts of political polarisation, they find that elections affect the types of treaties that leaders are willing to sign, focussing on the resulting bargaining positions which influence potential renegotiations after the election. They compare how *hawkish* or *dovish* incumbents manage the tradeoff between electoral prospects and policy success and conclude that the former seek to differentiate themselves from their challenger by rejecting deals they would myopically accept, while for the latter, maximising electoral prospects paradoxically leads to

them cutting better deals.

Some aspects of the question at hand have also been investigated empirically. Cazals and Sauquet (2015) analyse the effect that upcoming domestic elections have on the ratification timing of an IEA. They find that costly ratification of IEAs tends to be delayed to post-electoral periods. However, in developing countries, where the cost of ratification often tends to be lower, ratification may give rise to indirect advantages (e.g., foreign assistance) which make them prone to ratify pre-electorally. Furthermore, List and Sturm (2006) consider environmental policy of US state senators prior to elections and make use of the fact that binding term limits exist. Therefore, they are able to analyse whether environmental spending systematically varies depending on whether a governor is up for reelection or not. They find that it does, and that the direction of the policy distortion depends on the size of the environmental lobby in a country: in "green" states, governors advance less environmentally friendly policies once their term is up, and vice versa in "brown" states. Also, they show that the lower the political competition in a state, the less environmental policy is manipulated.

This paper builds on and complements model aspects of both Köke and Lange (2017) and Battaglini and Harstad (2020) while it aims at answering a question closely related to Buisseret and Bernhardt (2018) and Melnick and Smith (2022). My model enriches the bilateral treaty setup by Battaglini and Harstad (2020) along two main dimensions: Firstly, in my setup the environmental externality goes both ways and therefore makes country 2 not completely passive, in the sense that they have to agree with the treaty. Secondly, the setup is generally more nuanced, in that emission choices are modelled specifically as a result of reduction pledges, and that it allows for non-ratification (as opposed to compliance vs. non-compliance). In combination, this leads to the fact that on top of being able to replicate the results from Battaglini and Harstad (2020), additional and different equilibrium outcomes can be observed. Constrasting Melnick and Smith (2022), whose model endows the foreign country with agenda-setting power and considers a treaty merely about cost-sharing without a public goods characteristic, my model incorporates continuous policy choice, while abstracting from renegotiation. On top of that, all of the aforementioned papers largely abstract from political polarisation and its impacts on the degree to which international treaties are influenced.

2.3 The Model

We consider two countries, country 1 and 2, which negotiate a bilateral environmental agreement on the levels of GHG emissions. In the status quo, both countries choose emission levels such as to maximise the domestic welfare levels of whoever is in government. By doing so, due to the fact that GHG emissions are a transnational pollutant, countries do not take into account the negative externality their emission choice causes to the other country. The goal of the treaty is to commit to emission reductions relative to the status quo, and it depends on (i) who sits at the negotiation table and (ii) the political agenda put forward by the parties. We will analyse a situation in which a domestic election takes place in country 1 prior to the ratification of the treaty. It therefore is possible that the incumbent party, who

negotiates the treaty, will be replaced in the election and that therefore the challenging party will decide on the ratification of the agreement. Note, however, the challenger cannot renegotiate the treaty after an election win. In the case of non-ratification, emission levels are chosen non-cooperatively.

Countries are a priori, that is in the absence of domestic political competition, identical in terms of their environmental preferences. This allows us the isolate the effects of domestic political competition from potential effects of heterogeneity between the two countries. Regarding negotiation power, country 2 is modelled as a passive counterpart to country 1, as they do not take an active role in the design of the treaty but rather just accept or reject its contents. More precisely this means that the incumbent government of country 1 suggests a *take-it-or-leave-it* offer, which is restricted by country 2's participation constraint.

Note that generally in a political economy context, the socially optimal allocation depends on the distribution of environmental preferences across the varying actors involved. Therefore, without making a distributional assumption, social welfare implications cannot be discussed. As a benchmark, we will thus consider the optimal treaty choice for the median voter, as well as the treaty choice of the party in power if they did not face election pressure.

2.3.1 Agency structure

In each country, domestic emissions e_i lead to benefits from productive activities according to a concave quadratic benefit function $B(e_i)$, while global emissions cause linear environmental damages $D(E)$ with $E = e_1 + e_2$:

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

where ϵ denotes the business-as-usual emissions, capturing the emission level if the economy ran at full capacity and no emission reductions were beneficial. The parameter α measures carbon efficiency, that is, how much GDP a country can produce per unit of domestic emissions and β indicates the level of environmental damage in monetary terms caused per unit of global emissions.

The environmental preference parameter θ captures an agent's valuation of environmental damage costs. We assume that in the absence of political economy considerations, the two countries' median voters share the same value of θ. We will focus on political competition within country 1 and thus assume that country 2 is represented by a government with median voter preferences. We therefore normalise $\theta_2 = \theta_{2,M} = \theta_{1,M} = 1$, where subscript M stands for median voter.

Within country 1, there are two competing parties: the party in power at the start of the game is the incumbent i, the other is the challenger j in the upcoming election. We assume that one party is "greener" (G) than the median voter and the other is "browner" (B), that is, they either have a higher or lower willingness to pay for environmental damage reduction than the median voter. Consequently, it holds

that $\theta_{1,B} \leq \theta_{1,M} \leq \theta_{1,G}$. More precisely, the preference distance to the median voter is captured by the *polarisation* parameter ϕ:

$$\theta_{1,M} = 1, \qquad \theta_{1,G} = 1 + \phi, \qquad \theta_{1,B} = 1 - \phi. \tag{2}$$

Domestic welfare for any given agent $k \in \{M, B, G\}$ is defined by the difference between benefit and damage function, whereas the damage function is weighted with the agent's respective preference parameter θ:

$$W_{1,k}(e_1, e) = B(e_1) - \theta_{1,k} D(E), \tag{3}$$

$$W_2(e_2, e) = B(e_2) - D(E). \tag{4}$$

It follows that for a given value of global emissions, the welfare level of a greener agent (that is, with a higher value of θ) is always lower than that of a browner agent.

2.3.2 Treaty on Emission Reductions

The environmental treaty is modelled as the two countries cooperating with respect to emissions reductions. We consider an agreement design in which emissions are reduced proportionally to the status quo level, that is the non-cooperative Nash equilibrium resulting from incumbents in both countries maximising domestic welfare as given by (3) and (4), taking the emission choice of the other country as given. Note that country 1's status quo depends on the incumbent's preference parameter $\theta_{1,i}$ (henceforth I will use the simplified notation θ_i):

$$e_{1,i}^{\text{sq}} = \epsilon - \frac{\beta}{\alpha}\theta_i, \qquad e_2^{\text{sq}} = \epsilon - \frac{\beta}{\alpha}. \tag{5}$$

When negotiating a treaty, the incumbent government $i \in \{B, G\}$ of country 1 suggests a treaty parameter $\delta_i \in [0, 1]$, specifying the amount of emission reduction in the agreement. There exists an *preferred* value for the treaty parameter in the absence of an election from the perspective of country 1's incumbent. However, i might want to suggest a different value, taking into account the upcoming election. Due to the fact that country 2 does not have any negotiation power, they are assumed to ratify any treaty that makes them at least indifferent to the non-cooperative outcome, that is, the outcome in the absence of an agreement. Note that i suggests a single parameter, meaning that the two countries both reduce their emissions to an equal proportion. More formally, the incumbent i suggests a value δ_i, which determines agreement emission levels as follows:

$$\tilde{e}_{1,i} = \delta_i e_{1,i}^{\text{sq}}, \qquad \tilde{e}_2 = \delta_i e_2^{\text{sq}}.$$

The optimal treaty parameter in the absence of an election $\hat{\delta}_i$, henceforth called the *no-election treaty parameter*, is given as a solution to:

$$\max_{\delta_i \in [0,1]} W_i(\delta_i, \theta_i) = B\left(\tilde{e}_{1,i}(\delta_i)\right) - \theta_i D\left(\tilde{e}_{1,i}(\delta_i), \tilde{e}_2(\delta_i)\right)$$
$$\Rightarrow \quad \hat{\delta}_i(\theta_i) = \frac{1 + \beta\theta_i(\beta + \beta\theta_i - 3)}{(1 - \beta\theta_i)^2}. \tag{6}$$

Note that the median voter, having differing environmental preferences than the incumbent, has a different optimal value for δ_i, which is given as follows:

$$\max_{\delta \in [0,1]} W_M(\delta, \theta_i) = B(\tilde{e}_{1,i}(\delta)) - \theta_M D(\tilde{e}_{1,i}(\delta), \tilde{e}_2(\delta))$$
$$\Rightarrow \quad \delta_M^*(\theta_i) = \frac{1 + \beta(\beta(1 + \theta_i) - 2 - \theta_i)}{(1 - \beta\theta_i)^2}. \tag{7}$$

If the incumbent is green $\hat{\delta}_i < \delta_M^*$, and vice versa for a brown incumbent. Note that (7) does not indicate the median voter's optimal treaty choice in general, but the optimal treaty under a specific incumbent party, since it relates to i's status quo emissions (and not their own).

2.3.3 Timing

The game is formulated as a four stage game, with the timing given as follows:

1. **Agreement Stage**
   The two countries negotiate a bilateral agreement as described in Section 2.3.2. The treaty is then characterised by the parameter δ_i which maps into corresponding emission levels $\tilde{e}_1, \tilde{e}_2$ as defined by (10) and (11).

2. **Election Stage**
   An election takes place in country 1, where the median voter compares their welfare under the two parties. The reelection probability for the incumbent also stochastically depends on a relative popularity shock, as stated in Section 2.4.3.

3. **Ratification Stage**
   The election winner decides whether to ratify the agreement negotiated in Stage 1, which gives rise to ratification intervals of δ for both the incumbent and the challenger as detailed in Section 2.4.2.

4. **Emission Choice Stage**
   Domestic emission levels are chosen as a consequence of the ratification decision in Stage 3. In case of ratification, countries choose levels $\tilde{e}_1, \tilde{e}_2$, otherwise they set the non-cooperative emission levels $\hat{e}_1, \hat{e}_2$ as defined by (8) and (9).

2.4 Solving the Model

The game is solved by backwards induction and we are looking for subgame perfect Nash equilibria. Hence, in this section equilibrium outcomes of each stage will be discussed depending on which party serves as the incumbent and on the degree of political polarisation.

2.4.1 Emission Choice Stage

In the emission choice stage, the government elected in Stage 2 chooses their country's emission level depending on whether they opted to ratify the agreement in Stage 3.

The alternative to no ratified treaty is the non-cooperative outcome. This means that country 2 and the elected government (which is either the incumbent i or the challenger j) in country 1 maximise domestic welfare (3) and (4) resulting in the following non-cooperative emission levels:

$$\hat{e}_1(\theta_h) = \epsilon - \frac{\beta}{\alpha}\theta_h, \qquad h = i, j\,, \tag{8}$$

$$\hat{e}_2 = \epsilon - \frac{\beta}{\alpha}. \tag{9}$$

In case the treaty is ratified in Stage 3, the election winner sets emission levels such as to comply with the agreement negotiated in Stage 1, that is, to reduce status quo emissions according to δ_i. Emission levels are then given by:

$$\tilde{e}_1(\delta_i, \theta_i) = \delta_i \left(\epsilon - \frac{\beta}{\alpha}\theta_i\right), \tag{10}$$

$$\tilde{e}_2(\delta_i) = \delta_i \left(\epsilon - \frac{\beta}{\alpha}\right). \tag{11}$$

Importantly, note that if the challenger j wins the election and chooses to ratify the agreement, emission reductions will relate to the status quo emissions $e^{\text{sq}}_{1,i}$, that is, the level of $\tilde{e}_1$ is assumed to be fixed after Stage 1. This assumption seems intuitive thinking about a brown challenger who wins the election: if they choose to ratify, they adhere to the terms of the agreement as negotiated by the green incumbent, which results in lower emissions than what they would choose in the absence of a treaty. However, the assumption can seem problematic in the case of a green challenger who wins the election: if parties are highly polarised, it is possible that the green party's non-cooperative outcome is more ambitious than the negotiated treaty. In this case, the green challenger would, under a ratified treaty, reduce emissions by less than what they would do non-cooperatively. A possible argument in favour of this assumption could be that the treaty negotiation essentially captures the setting up of an international permit market, where the negotiating party decides on the permit supply, which for some time horizon after the election is fixed. Another argument could be that if the green challenger ratifies a treaty too

weak from their perspective, they would potentially forego the opportunity to negotiate a new and more ambitious treaty in the near future. Of course, however, this line of argument lies outside this model framework. In Section 2.6.1 I discuss an extension, in which the assumption of binding treaty emission is dropped and they merely serve as an upper bound which can be freely undercut by the party in power.

Without loss of generality, we can normalize the problem by setting $\alpha = \epsilon = 1$. Also, throughout the paper, the range $\beta \in [0, 0.15]$ for the marginal damage parameter will be assumed. This is a very non-restrictive assumption, since $\beta = 0.15$, from a median voter perspective, would imply the worst case scenario of unmitigated climate change corresponding to approximately a 45% decrease in global GDP. While this is certainly much higher than what is commonly found to be realistic with respect to climate change in a global context (see, e.g., Hänsel et al. 2020), allowing for such high values might make sense from a more local perspective, for example for countries in Southeast Asia, where local impacts are expected to be significantly higher than the global average (see, e.g., Swiss Re 2021). In any case, allowing for such high values of β ensures that results are not driven by a too optimistic evaluation of environmental damages.

2.4.2 Ratification Stage

Whoever is in charge at the ratification stage, that is, the election winner in Stage 2, will decide between ratifying the treaty on the table or not. As previously stated, we assume that renegotiation of the treaty is not available to the governing party after the election. This is a reasonable assumption in the context of this model, since historically it can be argued that treaty negotiations usually take place over a long time horizon and that it is not possible for a newly elected government of one country to immediately renegotiate an international agreement, as seen, for example, in the case of the US and the Kyoto Protocol. This feature of the model contrasts the setup of Buisseret and Bernhardt (2018) as well as Melnick and Smith (2022), where agreements made before an election serve as a starting point for any subsequent renegotiation.

Ratification Intervals

The incumbent i or the challenger j, when elected, ratify the agreement whenever:

$$\Delta W_h = W_h(\tilde{e}_1(\delta_i, \theta_i), \tilde{e}) - W_h(\hat{e}_1(\theta_h), \hat{e}) \geq 0, \qquad h = i, j\,. \tag{12}$$

This leads to two intervals for which elected party will ratify the agreement, each defined by the two

threshold values for δ:

$$[\underline{\delta}^i, \bar{\delta}^i] = \left[\max\left\{0, \frac{1+\beta\theta_i\left(\beta(2+\theta_i)-4\right)}{(\beta\theta_i-1)^2}\right\}, 1\right], \tag{13}$$

$$[\underline{\delta}^j, \bar{\delta}^j] = \left[\max\left\{0, \frac{1-\beta\left[\theta_i-\theta_j(\beta+\beta\theta_i-2)\right]-\sqrt{M}}{(\beta\theta_i-1)^2}\right\},\right.$$
$$\left.\min\left\{\frac{1-\beta\left[\theta_i-\theta_j(\beta+\beta\theta_i-2)\right]+\sqrt{M}}{(\beta\theta_i-1)^2}, 1\right\}\right]. \tag{14}$$

where $M = \beta^2\theta_j(\beta-1)\left[\theta_j(\beta+2\beta\theta_i-3)-2\theta_i(\beta\theta_i-1)\right]$. Also, note that the incumbent's upper threshold value corresponds to no emission reductions, that is, to the non-cooperative outcome.

We will see in the following that it depends on which party is the incumbent and which is the challenger in order to state how these thresholds relate to each other. We define as follows:

$$\Delta\underline{\delta} \equiv \underline{\delta}^i - \underline{\delta}^j, \tag{15}$$

$$\Delta\bar{\delta} \equiv \bar{\delta}^i - \bar{\delta}^j. \tag{16}$$

While the incumbent of country 1 suggests a treaty parameter δ_i, country 2 is assumed to not have any negotiation power. Still, country 1's treaty suggestion is limited by country 2's participation constraint, that is, the fact that country 2 has to be better off than under no agreement, in which case they expect the incumbent's non-cooperative emission choice. The corresponding ratification thresholds for country 2 are then given as follows:

$$\Delta W_2 = W_2(\tilde{e}_2(\delta_i, \theta_i, \theta_2), \tilde{e}) - W_2(\hat{e}_2(\theta_2), \hat{e}) \geq 0 \tag{17}$$

$$\Leftrightarrow \quad [\underline{\delta}_2(\theta_i), \bar{\delta}_2(\theta_i)] = \left[\frac{1+\beta(2\beta\theta_i+\beta-4)}{(\beta-1)^2},\ 1\right]. \tag{18}$$

Country 2's ratification interval only depends on the incumbents' preference parameter. This is due to the fact that their treaty partner in the agreement stage is the incumbent and even if the challenger were to win the election and ratify the treaty, country 2's emission reduction commitment only relates to the treaty signed with the incumbent. Implictly, we assume country 2 to not be sophisticated enough to anticipate the possibility of facing the challenger's non-cooperative outcome if they were elected and did not ratify.

The threshold values (13) and (14) can then be ordered and will give rise to a partition of ranges for the treaty parameter. The ordering will depend on which party is the incumbent and the size of the polarisation parameter ϕ. Four different cases can emerge:

(A) $\delta \in [\underline{\delta}^i, \bar{\delta}^i], \delta \notin [\underline{\delta}^j, \bar{\delta}^j]$: only the incumbent ratifies,

(B) $\delta \notin [\underline{\delta}^i, \bar{\delta}^i], \delta \in [\underline{\delta}^j, \bar{\delta}^j]$: only the challenger ratifies,

(C) $\delta \in [\underline{\delta}^i, \bar{\delta}^i], \delta \in [\underline{\delta}^j, \bar{\delta}^j]$: both ratify,

(D) $\delta \notin [\underline{\delta}^i, \bar{\delta}^i], \delta \notin [\underline{\delta}^j, \bar{\delta}^j]$: none ratify.

Note that technically, there is an additional case in which a treaty that would be ratified by at least one of the parties in country 1 but not by country 2. However, from a theoretical point of view there is no difference to case *D* regarding the resulting emission choices. Thus, henceforth, this scenario will be captured by case *D*.

In the following, we will discuss specific ratification intervals in the context of a green and a brown incumbent.

Green incumbent

In a first step, let us assume the green party serves as the incumbent. We can therefore assign preference parameters $\theta_i = 1 + \phi$ and $\theta_j = 1 - \phi$. Note that if the parties are too polarised, the challenger would never ratify a treaty suggested by the incumbent. This is the case if no real value δ_i satisfies (12).

Lemma 1 (Existence of Ratification Interval for Brown Challenger)
The challenger's ratification thresholds exist if it holds that:

$$\phi \leq \bar{\phi}(\beta) = \frac{5\beta - 5 + \sqrt{(\beta - 1)(41\beta - 25)}}{8\beta}. \tag{19}$$

In addition it holds that:

$$\frac{d\bar{\phi}}{d\beta} > 0. \tag{20}$$

The relation (20) states that for higher environmental damages, a higher degree of polarisation allow for a ratification interval by the brown challenger.

We can now specify ratification intervals (13) and (14) in case of a green incumbent as given in Proposition 12.

Proposition 12 (Stage 3: Ratification Intervals with $i = G$)
In the case of a green incumbent, the incumbent's and country 2's ratification thresholds are given as follows:

$$[\underline{\delta}^{i=G}, \bar{\delta}^{i=G}] = \left[\max\left\{0, \frac{1 + \beta(1+\phi)\left[\beta(3+\phi) - 4\right]}{(\beta(1+\phi) - 1)^2}\right\}, 1\right], \tag{21}$$

$$[\underline{\delta}_2(\theta_G), \bar{\delta}_2(\theta_G)] = \left[\frac{1 + \beta[\beta(3+2\phi) - 4]}{(\beta - 1)^2}, 1\right]. \tag{22}$$

The challenger's ratification thresholds exist when $\phi \leq \bar{\phi}$. In that case, they are given by:

$$[\underline{\delta}^{j=B}, \bar{\delta}^{j=B}] = \left[\frac{1+\beta\left[\phi-3\right]-\beta^2\left[\phi^2+\phi-2\right]-\sqrt{M_{j=B}}}{(\beta(1+\phi)-1)^2}, \right. \\ \left. \frac{1+\beta\left[\phi-3\right]-\beta^2\left[\phi^2+\phi-2\right]+\sqrt{M_{j=B}}}{(\beta(1+\phi)-1)^2} \right], \tag{23}$$

where $M_{j=B} = \beta^2(\beta-1)(\phi-1)\left[1-5\phi+\beta(4\phi^2+5\phi-1)\right]$.

Proposition 13 (Stage 3: Comparative Statics with $i = G$)
The following conditions hold for the equilibrium ratification intervals under the condition that thresholds exist and that they are within the interval $[0, 1]$*:*

$$\frac{d\underline{\delta}^{i=G}}{d\phi} < 0, \qquad \frac{d\bar{\delta}^{i=G}}{d\phi} = 0, \qquad \frac{d\underline{\delta}^{j=B}}{d\phi} > 0, \qquad \frac{d\bar{\delta}^{j=B}}{d\phi} < 0.$$

Proposition 13 states that while the incumbent's upper threshold is independent of ϕ, their lower ratification threshold decreases with the distance from the median voter. Intuitively this means that a greener incumbent will be more willing to ratify strict treaties. For the brown challenger, the upper threshold decreases and the lower threshold increases in ϕ, meaning that their ratification interval becomes more narrow with increasing polarisation. Consequently, higher polarisation leads to a more narrow range of δ_i that would allow for ratification by both parties.

Proposition 14 (Ordering of Parties' Ratification Thresholds with $i = G$)
In the case of the green incumbent and whenever $\phi \leq \bar{\phi}$, the ratification thresholds (21) *and* (23) *relate to each other as given in the following:*

$$\Delta\underline{\delta}^{i=G} \leq 0, \tag{24}$$

$$\Delta\bar{\delta}^{i=G} \geq 0. \tag{25}$$

Proposition 14 states that the green incumbent will always sign stricter treaties than the brown challenger. While no treaty is unambitious enough for the incumbent (technically they can negotiate a treaty with $\delta_i = 1$, which corresponds to their non-cooperative outcome), the challenger will not always sign such treaties. The reason for this is that if emission reductions are negligible, the positive effect of damage reductions (also via less externalities from country 2) does not outweigh the negative effect of not being able to choose emissions freely according to their own optimal non-cooperative outcome.

Lemma 2 (Ordering of Countries' Ratification Thresholds with $i = G$)
The lower ratification threshold of parties in country 1 relate to that of country 2 as follows:

$$\underline{\delta}^{i=G} - \underline{\delta}_2 \leq 0, \tag{26}$$

$$\underline{\delta}^{j=B} - \underline{\delta}_2 \geq 0. \tag{27}$$

Lemma 2 states that country 1 with a green incumbent is willing to ratify more strict treaties than allowed for by country 2's participation constraint. This implies that in some cases, country 2's lower ratification threshold is binding, rather than that of the green incumbent. The brown challenger, conversely, ratifies less ambitious treaties than country 2.

This leads us to a discussion of which cases arise along the spectrum of possible treaty parameters, following Proposition 14 and Lemma 2:

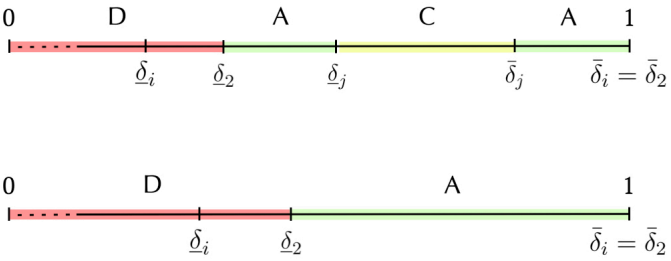


Figure 6: Green incumbent ratification thresholds

The two scenarios are distinguished by whether ratification thresholds for the challenger exist, following the threshold value $\bar{\phi}$ as defined by Lemma 1. If the thresholds for the challenger do not exist, the feasible range for the treaty parameter is limited by country 2. Note that the colours in Figure 1 and henceforth indicate which parties ratify the treaty: green and brown for the green and brown party respectively, yellow for both parties and red for none.

Brown incumbent

Next, we consider the scenario in which the brown party is the incumbent and therefore environmental preference parameters are given by $\theta_i = 1 - \phi$ and $\theta_j = 1 + \phi$. Ratification intervals following (13) and (14) are then as stated in Proposition 15.

Proposition 15 (Stage 3: Ratification Intervals with $i = B$)
In the case of a brown incumbent, ratification thresholds are given as follows:

$$[\underline{\delta}^{i=B}, \bar{\delta}^{i=B}] = \left[\frac{1+\beta(\phi-1)\left[4+\beta(\phi-3)\right]}{\left(1+\beta(\phi-1)\right)^2}, 1\right], \tag{28}$$

$$[\underline{\delta}_2(\theta_B), \bar{\delta}_2(\theta_B)] = \left[\frac{1+\beta(\beta(3-2\phi)-4)}{(\beta-1)^2}, 1\right]. \tag{29}$$

The challenger's ratification thresholds always exist and are given by:

$$[\underline{\delta}^{j=G}, \bar{\delta}^{j=G}] = \left[\max\left\{0, \frac{1-\beta\left[3+\phi+\beta(\phi-2)(1+\phi)\right]-\sqrt{M_{j=G}}}{\left(1+\beta(\phi-1)\right)^2}\right\}, \right. \\ \left. \frac{1-\beta\left[3+\phi+\beta(\phi-2)(1+\phi)\right]+\sqrt{M_{j=G}}}{\left(1+\beta(\phi-1)\right)^2}\right\}\Bigg], \tag{30}$$

with $M_{j=G} = \beta^2(1-\beta)(1+\phi)\left(1+5\phi+\beta\left[\phi(4\phi-5)-1\right]\right)$.

Proposition 16 (Stage 3: Comparative Statics with $i = B$)
The following conditions hold for the equilibrium ratification intervals under the condition that they are within the interval $[0, 1]$*:*

$$\frac{d\underline{\delta}^{i=B}}{d\phi} > 0, \qquad \frac{d\bar{\delta}^{i=B}}{d\phi} = 0, \qquad \frac{d\underline{\delta}^{j=G}}{d\phi} < 0, \qquad \frac{d\bar{\delta}^{j=G}}{d\phi} < 0.$$

Proposition 16 states that the lower ratification threshold for the brown incumbent increases in the distance from the median voter, whereas the upper threshold is independent of ϕ. Intuitively, the browner the incumbent, the less strict the treaty can be for them to ratify. For the challenger, higher polarisation decreases both upper and lower thresholds. The greener the challenger, the more ambitious the treaty can be on the lower end and has to be on the upper end, for them to ratify. If a treaty is too weak, the damage reduction does not compensate for insufficiently low emission levels. This essentially means that their ratification interval shifts downwards.

Proposition 17 (Ordering of Parties' Ratification Thresholds with $i = B$)
In the case of the brown incumbent, the ratification thresholds (28) *and* (30) *relate to each other as given in the following:*

$$\Delta\underline{\delta}^{i=B} \geq 0, \tag{31}$$

$$\Delta\bar{\delta}^{i=B} \geq 0. \tag{32}$$

Proposition 17 states that the green challenger will sign more ambitious treaties than the brown in-

cumbent. However, at the upper end of the spectrum, there are very unambitious treaties that the incumbent will sign and the challenger not. Intuitively, these are contracts that are so unambitious in terms of damage reduction that the green challenger is better off with their non-cooperative outcome.

Lemma 3 (Ordering of Countries' Ratification Thresholds with $i = B$)
The incumbent's lower ratification threshold relates to that of country 2 as follows:

$$\underline{\delta}^{i=B} - \underline{\delta}_2 \geq 0, \tag{33}$$
$$\underline{\delta}^{j=G} - \underline{\delta}_2 \leq 0. \tag{34}$$

Lemma 3 states that country 1 with a brown incumbent is willing to ratify less ambitious treaties than would be allowed for by country 2's participation constraint, and vice versa for the green challenger.

Consequently, the following cases arise, as postulated by Proposition 17 and Lemma 3:

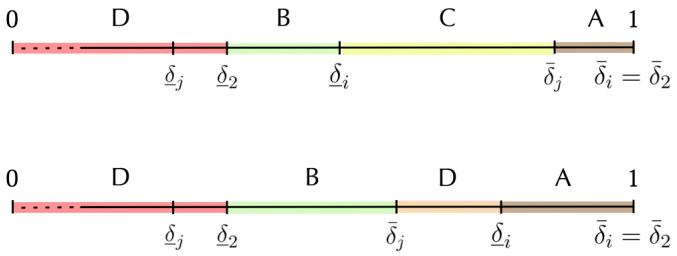


Figure 7: Brown incumbent ratification thresholds

The two scenarios are separated by whether ratification intervals of the incumbent and challenger overlap or not, that is, depending on $\underline{\delta}_i \lessgtr \bar{\delta}_j$. If they do not overlap, this means that no treaty parameter leads to ratification by both parties, as depicted in the second scenario (no area C).

Lemma 4 (Ordering of Countries' Ratification Intervals with $i = B$)
There exists a threshold value $\tilde{\phi}$ for ratification intervals to touch, i.e. at which $\underline{\delta}_i = \bar{\delta}_j$. Then, if:

$$\phi \leq \tilde{\phi}(\beta) \ \in [0.768, 0.8). \tag{35}$$

a common ratification interval exists.

Hence, if polarisation is very high, there exists an interval for the treaty parameter, which will not be signed by any of the two parties, since it is too ambitious for the brown incumbent while being not ambitious enough for the green challenger.

2.4.3 Election Stage

The domestic election is modelled as devised by Battaglini and Harstad (2020). The median voter faces the choice between the incumbent i and the challenger j and considers how each will affect their welfare: depending on the election outcome, country 1 can either (i) be part of the treaty and choose emissions as negotiated by i or (ii) live in a world without a treaty where both countries choose the non-cooperative outcome. In the former case, both parties will act identically whereas in the latter, the outside option differs between the two. The median voter can, as defined by the ratification intervals derived in Section 2.4.2, anticipate what the consequence of electing either of the two parties is.

The median voter's welfare difference between a government i and j is denoted by ΔW_M and the incumbent is consequently re-elected whenever:

$$\Delta W_M \equiv W_M^i - W_M^j \geq \Omega, \qquad \text{where } \Omega \sim U\left[-\frac{1-z}{\sigma}, \frac{z}{\sigma}\right],$$

which gives rise to the reelection probability for the incumbent party:

$$p^l(\delta_i) = \sigma \Delta W_M^l + z, \qquad l \in \{A, B, C, D\}. \tag{36}$$

This reelection probability is a function of the treaty parameter δ_i that determines which case A – D emerges. Therefore, the reelection probability between cases differs, since W_M^i and W_M^j depend on whether ratification occurs in Stage 3.

The parameter $z \geq 0.5$ quantifies an incumbency advantage, that is, the reelection probability for the incumbent in the absence of any policy differences between the two parties. The parameter σ captures the density of a popularity shock. A high value of σ (low variance) means that policy differences are more likely to dictate the outcome of the election, whereas low values of σ (high variance) increase noise and thus make random popularity shocks more important. The parameter can therefore also be interpreted as a value for policy salience. An example of such a shock could be an exogenous change in the political climate with respect to environmental issues, as for example, the Fukushima nuclear disaster in 2011.

Lemma 5 (Restrictions on Shock Density)
For reelection probabilities to be interior in $(0, 1)$*, the variance in the popularity shock is restricted to:*

$$\sigma < \min\left\{\frac{1-z}{\Delta W_M^l}, \frac{z}{|\Delta W_M^l|}\right\}, \tag{37}$$

which is most restrictive for the case $l \in \{A, B, C, D\}$ *for which* $|\Delta W_M^l|$ *is highest.*

This leads to the full characterisation of reelection probabilities in Stage 2.

Proposition 18 (Stage 2: Reelection Probabilities)
Given that $\sigma \in (\underline{\sigma}, \bar{\sigma})$ and $z \geq 0.5$, reelection probabilities for cases A – D are defined as follows:

(A) Only the incumbent party ratifies

$$\begin{aligned} \Delta W_M^A &= W_M(\tilde{e}_1(\theta_i, \delta_i), \tilde{e}) - W_M(\hat{e}_1(\theta_j), \hat{e}) \\ &\Rightarrow \; p^A(\delta_i) = \sigma \Delta W_M^A + z. \end{aligned} \tag{38}$$

(B) Only the challenging party ratifies

$$\begin{aligned} \Delta W_M^B &= W_M(\hat{e}_1(\theta_i), \tilde{e}) - W_M(\tilde{e}_1(\theta_i, \delta_i), \hat{e}) \\ &\Rightarrow \; p^B(\delta_i) = \sigma \Delta W_M^B + z. \end{aligned} \tag{39}$$

(C) Both parties ratify

$$\begin{aligned} \Delta W_M^C &= 0 \\ &\Rightarrow \; p^C = z. \end{aligned} \tag{40}$$

(D) None of the parties ratify

$$\begin{aligned} \Delta W_M^D &= W_M(\hat{e}_1(\theta_i), \hat{e}) - W_M(\hat{e}_1(\theta_j), \hat{e}) \\ &\Rightarrow \; p^D = \sigma \Delta W_M^D + z. \end{aligned} \tag{41}$$

Straightforwardly, the median voter's welfare level is affected in the cases where the incumbent and the challenger will take different ratification decision (A and B). In the case where both parties will ratify C, this is not the case because the challenger is tied to the treaty negotiated by the incumbent. In the last case D, again the median voter's welfare levels are different since the two parties will choose differing non-cooperative emission levels. Note that the reelection probability is a function of the treaty parameter in cases A and B but not in cases C and D.

2.4.4 Agreement Stage

The incumbent government negotiates an agreement such that their expected welfare is maximised:

$$\max_{\delta_i} \; p(\delta_i)\Big[W_i(\text{'i in power'}) + R\Big] + (1 - p(\delta_i))\Big[W_i(\text{'j in power'})\Big], \tag{42}$$

and $W_i(\cdot)$ and $p(\delta_i)$ depend on cases A – D, as detailed in the previous sections.

R denotes the rent from staying in office, which can capture any inherent benefits from staying in power. Battaglini and Harstad (2020) refer to R as an additional measure for political polarisation: the further apart the two parties, the more important holding the office is, for example, to influence domestic policy unrelated to emission choice. Furthermore, the level of office rents might differ between political systems: presidential systems would then be associated with higher values of R, as opposed to parliamentary systems, in which the surplus from being in office is more spread across political actors and being in office comes with less power to push one's own agenda.

The incumbent government chooses δ_i, perfectly anticipating in which case they will end up according to the derived ratification intervals. Therefore, they compute expected maximum welfare levels for each case and then opt for the case with the highest value. Note that whenever both parties would ratify the agreement, that is, case C, it is welfare-maximising for the incumbent to set $\tilde{\theta} = \theta_i$. In that case the objective function is given by:

$$\begin{aligned} W_i^C &= p^C \Big[W_i(\tilde{e}_1(\theta_i, \delta_i)) + R \Big] + (1 - p^C) \Big[W_i(\tilde{e}_1(\theta_i, \delta_i)) \Big] \\ &= W_i(\tilde{e}_1(\theta_i, \delta_i)) + zR, \end{aligned} \tag{43}$$

which qualitatively corresponds to the maximisation problem of (6). Intuitively, since the reelection probability is not influenced by the choice of agreement, we therefore see no distortion of policy choice in this case, which is why we sometimes end up in a *first-best* outcome. However, we will see that choosing $\hat{\delta}_i$ does not necessarily lead to case C, as the challenger might not want to ratify this treaty.

Depending on the level of political polarisation, the median voter prefers a treaty that is either signed by both parties, henceforth referred to as a *consensus treaty*, or in case of high polarisation, a treaty ratified only by the green party. Intuitively, if parties are sufficiently similar, a consensus treaty can achieve enough from the point of view of the median voter. However, if preferences are too different, the median voter favours a treaty strict enough to keep the brown party from ratifying over a watered-down consensus. The threshold value of polarisation which separates those two cases will depend on the incumbent party.

Note that the optimal treaty choice for the incumbent is presented in numerical examples in order to illustrate which different outcomes can occur. The following parameters will be assumed throughout:

$$z = 0.55, \quad \beta = 0.05, \quad \sigma = 0.8.$$

All of these parameter values are unexceptional, in the sense that they are not driving any of the results presented, and postulate (i) a 5 percentage point incumbency advantage, which following, for example, Gelman and King (1990) and Levitt and Wolfram (1997) are a *middle-of-the-road* estimate, (ii) environmental damages in the absence of climate change mitigation would constitute an approximately 18% reduction of GDP, and (iii) a shock density parameter to mirror environmental policy being

relatively salient such that 80% of policy differences between the contenders transmit into reelection probabilities.

In the figures of this section the incumbent's expected welfare level W_i will be plotted against the agreement parameter δ_i and the different colours of the functions will indicate the case $A - D$ the respective value of δ_i would give rise to. The dotted line shows the value of $\hat{\delta}_i$ as given by (6), that is the preferred treaty parameter in the absence of an election, and the pink line indicates the optimal parameter choice as a solution to (42). The blue line depicts the preferred treaty parameter for the median voter (relating to the incumbent's non-cooperative emissions) as given by (7).

Green incumbent

In this section, we discuss possible outcomes when the incumbent party is green, depending on the degree of political polarisation. Ratification thresholds and thus the partition of the δ-range into the four cases are given as depicted in Figure 1, and corresponding reelection probabilities are anticipated as given in Section 2.4.3.

The threshold value ϕ_G^M which determines whether the median voter prefers ratification by both (consensus treaty) or only the green incumbent (differentiation treaty) is implicitly defined by:

$$\delta_M^*(\theta_G, \phi_G^M) = \underline{\delta}_{j=B}(\phi_G^M). \tag{44}$$

For the parameter range assumed for β, $\phi_G^M(\beta) \in [0.1718, 0.1754]$. Whenever polarisation is below this value, the median voter prefers a consensus treaty, otherwise a differentiation treaty. More broadly, we define two ranges as low and high polarisation, as seen in Figure 8.

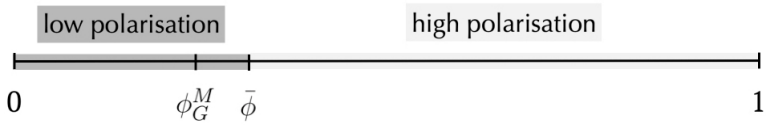


Figure 8: Polarisation ranges for $i = G$

Low polarisation

As depicted in Figure 8, polarisation is defined as low if a common ratification interval exists, that is if $\phi \leq \bar{\phi}$. In Figure 6 we can see that there are two areas in which case A arises, however, the incumbent only ever opts for treaties in the area with lower δ values or the area C.

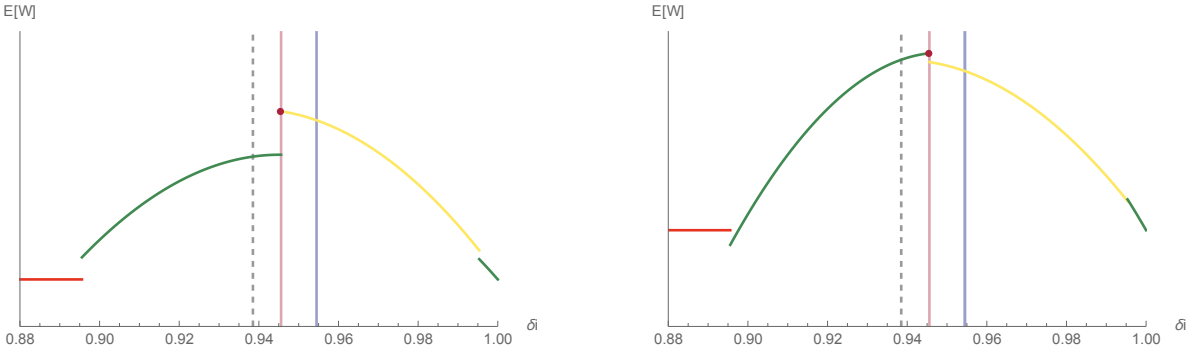
Proposition 19 (Treaty Outcomes with Low Polarisation for $i = G$)
For low levels of polarisation, that is for $\phi \leq \bar{\phi}$, it holds that:

1. *it is never optimal for the incumbent to choose a treaty in the upper area A, that is, $\delta \in [\bar{\delta}_j, 1]$,*

2. *the incumbent decides between a consensus and differentiation treaty (lower area A) depending on the size of the office rent R.*

We first consider the case in which the median voter prefers a consensus treaty, that is, whenever $\phi \leq \phi_G^M$. As stated in Proposition 19, the incumbent's treaty choice depends on the size of the office rent R, with the incumbent being indifferent between the two treaty types if $R = \bar{R}(\phi)$. Note that in both cases *C* and *A*, the optimal choice by the green incumbent is a treaty that is weaker than what they would prefer in the absence of an election.

Figure 9a illustrates a case in which the office rent is lower than the threshold value $\bar{R}$ and setting $\delta_i^* = \bar{\delta}_j + \epsilon$ is optimal. Reducing the ambition of the treaty by increasing δ versus $\hat{\delta}_i$ is profitable, as they prefer a less strict treaty which is ratified for sure over the risk of ending up with the challenger's non-cooperative outcome. Figure 9b contrastingly illustrates the case in which the incumbent is willing to take this risk by setting $\delta_i^* = \bar{\delta}_j - \epsilon$: reelection chances are increased by forcing a differentiation against the challenger, in which case the median voter prefers an agreement that is deemed too strict over the non-cooperative outcome by the brown challenger, that is $\Delta W_M^A > 0$ and hence $p^A(\delta_i^*) > p^C$. This increased reelection probability is, due to the office rent above $\bar{R}$ when the election is won, worth more than getting a desirable outcome in case of election loss.



(a) Consensus treaty with $\phi = 0.15, R = 0.5$ **(b)** Differentiation treaty with $\phi = 0.15, R = 1.25$

Figure 9: Treaty outcomes with low polarisation and $i = G$

There exists a narrow range $\phi \in (\phi_G^M, \bar{\phi})$ in which the median voter would prefer a differentiation treaty. From the incumbent's point of view the choice between a consensus and a differentiation treaty still depends on the size of the office rent in the same way as discussed. However, as opposed to the previous example, the median voter now achieves their preferred treaty only in case of a sufficiently high office rent $R > \bar{R}$.

Note that the first-best treaty, the case in which no distortion compared to the optimal choice in the absence of an election occurs, can materialise. This treaty is available for sufficiently low values of polarisation: for the range of the damage parameter assumed, this threshold value is approximately $\phi = 0.135$. Whether the incumbent eventually opts for this treaty still depends on the office rent.

High polarisation

With sufficiently high polarisation $\phi > \bar{\phi}$, the challenger never ratifies. Still, the incumbent has an incentive to depart from $\hat{\delta}_i$, as illustrated by Figure 10. In choosing the treaty parameter, the incumbent trades off the increased reelection probability, due to the fact that the chosen value is closer to the median voter's preferred value, and the lower welfare levels due to a less than desired treaty ambition.

Proposition 20 (Treaty Outcomes with High Polarisation for $i = G$)
For sufficiently high levels of polarisation, that is for $\phi > \bar{\phi}$, it is optimal for the incumbent to choose a treaty with $\delta_i^* \in (\hat{\delta}_i, \delta_M^*)$.

The higher the office rent, the more important it is for the incumbent to increase re-election probability and the more they assimilate towards the median voter. Thus, as R increases, $\delta_i^* \to \delta_M^*$.

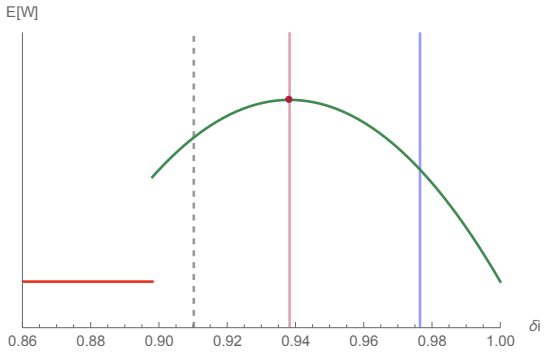


Figure 10: Assimilation treaty with $\phi = 0.6, R = 0.5$

Brown incumbent

Let us now assume that the incumbent party is brown and thus the ordering of ratification thresholds are as depicted in Figure 2 and reelection probabilities are anticipated to be as specified in Section 2.4.3. Again, we will discuss how different degrees of political polarisation will affect the optimal treaty choice.

The threshold value ϕ_B^M which determines whether the median voter prefers ratification by both parties or only by the green challenger is given as follows:

$$\phi_B^M(\beta) = \frac{3\beta - 3\sqrt{5\beta^2 - 14\beta + 9}}{2\beta} \tag{45}$$

For the parameter range assumed for β, $\phi_B^M(\beta) \in [0.327, 0.333]$. Whenever polarisation is below this value, the median voter prefers a consensus treaty, otherwise they would optimally get a treaty only ratified by the green challenger (case B). Figure 11 illustrates the three ranges of polarisation that arise.

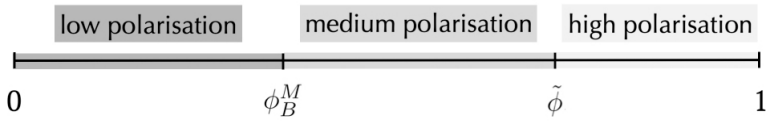


Figure 11: Polarisation ranges for $i = B$

Low & Medium polarisation

We define polarisation to be low if $\phi \leq \phi_B^M$, in which case the median voter prefers a consensus treaty. Polarisation is referred to as medium if $\phi_B^M < \phi \leq \tilde{\phi}$. This is the range in which the medium voter would want a differentiation treaty. While there always exists an overlap in ratification intervals of incumbent and challenger, depending on the size of the office rent, optimal treaty choices lie in the range of either case C or case A. As stated in Proposition 21, it is never optimal for the incumbent to choose a treaty that is only ratified by the challenger, even though that is what the median voter would prefer in some cases.

Proposition 21 (Treaty Outcomes with Low & Medium Polarisation for $i = B$)
For low & medium levels of polarisation, that is for $\phi \leq \tilde{\phi}$, it holds that:

1. *it is never optimal for the incumbent to choose a treaty of type B,*
2. *the incumbent decides between a consensus and differentiation treaty depending on the size of the office rent R,*
3. *for all values $\phi \leq \phi_B^M$, $\hat{\delta}_i$ is always within area C, making consensus treaties first-best.*

For low levels of polarisation, the optimal treaty choice is given by $\delta_i^* = \hat{\delta}_i$, which corresponds to no policy distortion. This treaty is also ratified by the green challenger, so that this is actually a special case of a consensus treaty in which the incumbent gets to choose their first-best treaty, as illustrated by Figure 12. Note that differentiation treaties are possible for very high levels of office rent, that is $R > 2.5$, the exact threshold depending on the polarisation parameter.

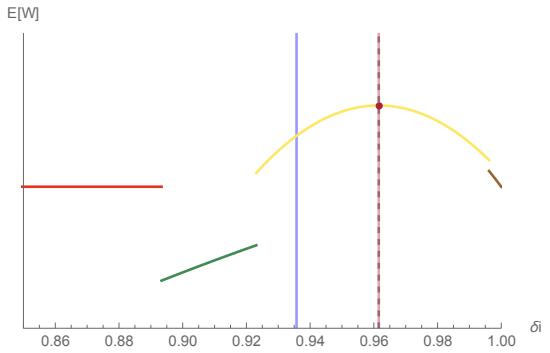
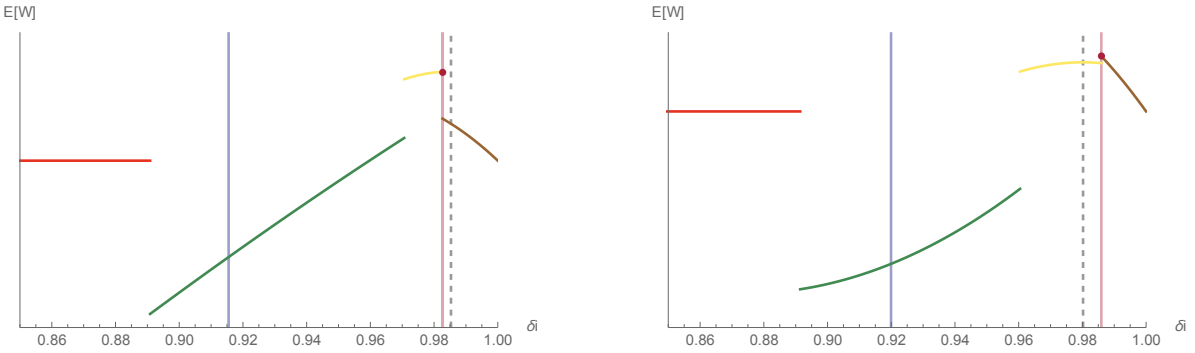


Figure 12: Treaty outcome with low polarisation $\phi = 0.25$ and $i = B$

In case of medium polarisation, the incumbent's decision between a differentiation and consensus treaty again depends on the size of the office rent, where the indifference point $\bar{R}$ is a function of polarisation ϕ. When $R < \bar{R}$ the incumbent optimally sets $\delta_i^* = \bar{\delta}_j + \epsilon < \hat{\delta}_i$, that is, just meets the challenger's ratification interval as pictured in Figure 13a. The true optimal treaty for the brown incumbent is not ambitious enough for the challenger to ratify and therefore the incumbent chooses to slightly decrease δ_i in order to ensure a ratified agreement even in case of election loss.



(a) Consensus treaty with $\phi = 0.7, R = 0.5$

(b) Differentiation treaty with $\phi = 0.6, R = 1.3$

Figure 13: Treaty outcomes with medium polarisation and $i = B$

Contrarily, if $R > \bar{R}$, the brown incumbent prefers to differentiate from the green challenger and optimally sets $\delta_i^* = \bar{\delta}_j + \epsilon > \hat{\delta}_i$ as seen in Figure 13b. By choosing a differentiation treaty (case A), the median voter is forced to compare the weak treaty to the non-cooperative outcome of the very green challenger, where $\Delta W_M^A > 0$ and thus $p^A(\delta_i^*) > p^C$. The high office rent then makes it worth for the incumbent to exploit this difference as opposed to choosing a better policy. Note that in both of the described cases, the median voter would prefer a treaty of type B, which never materialises.

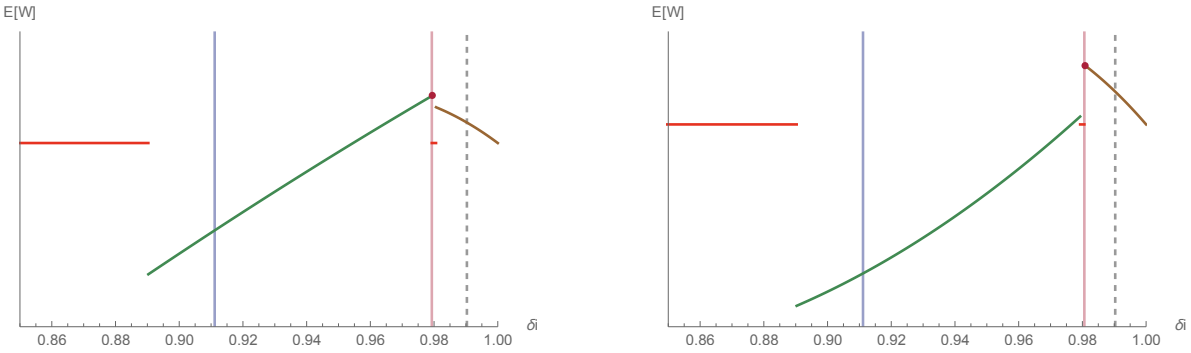
High polarisation

Lastly, polarisation is defined as high when no overlap in ratification intervals exists, that is whenever $\phi > \tilde{\phi}$. There even exists an interval $(\bar{\delta}_j, \underline{\delta}_i)$ which is not ratified by any of the two, as it is too ambitious for the incumbent and not ambitious enough for the challenger. However, as stated by Proposition 22, the incumbent never chooses such a treaty.

Proposition 22 (Treaty Outcomes with High Polarisation for $i = B$)
For high levels of polarisation, that is for $\phi > \tilde{\phi}$, it holds that:

1. *it is never optimal for the incumbent to choose a treaty of type D,*
2. *the incumbent decides between an insurance and differentiation treaty depending on the size of the office rent R.*

As illustrated in Figure 14, the choice between a so-called *insurance treaty* and differentiation treaty depending on the size of the office rent: if the office rent is sufficiently small, it can be optimal for the incumbent to suggest a treaty that they themselves would not ratify but their green challenger would, that is, $\delta_i^* = \bar{\delta}_j - \epsilon < \hat{\delta}_i$ as shown in Figure 14a. This strategy is beneficial and "low-risk" for the incumbent for two reasons: on the one hand, they are pursuing a "cheap" policy in case of reelection by not committing to any emission reductions. Given their low environmental preferences, the cost of country 2 not reducing emissions is relatively low. On the other hand, in case they are replaced, the negotiated treaty gives the the incumbent higher welfare levels than the challenger's non-cooperative policy choice would. Interestingly, $\Delta W_M^B(\delta_i^*) < 0$, meaning that the incumbent knowingly reduces their reelection probability. Therefore, as opposed to the differentiation treaty and due to the strong polarisation, they prioritise policy outcome over reelection. One can therefore interpret this choice as an insurance against a potential successor.



(a) Insurance treaty with $\phi = 0.8, R = 0.5$

(b) Differentiation treaty with $\phi = 0.8, R = 1$

Figure 14: Treaty outcomes with high polarisation and $i = B$

As seen in Figure 14b, if the office rent is sufficiently high, a differentiation treaty results because the focus is shifted towards increasing the reelection probability. However, in both cases, the resulting treaty is much too unambitious for the median voter and emission levels much higher than what they would prefer.

2.5 Numerical Illustration

This section provides a simple numerical illustration of the main model. This allows for a direct comparison between different treaty types in terms of emission levels and reelection probabilities. Also, a natural question arises regarding a ranking of the presented treaty types in terms of implications on welfare. Given the agency structure of the model, this is not an obvious exercise. The approach chosen here focusses on the median voter and how the treaty types compare to their optimal treaty under a

given administration as stated in (7). The welfare measure used is given in the following:

$$\Delta W_M^{loss} = \frac{W_M(\delta_M^*(\theta_i)) - \mathbb{E}[W_M(\delta_i^*(\theta_i))]}{W_M(\delta_M^*(\theta_i))} \tag{46}$$

Intuitively, this captures the percentage loss in median voter welfare that arises due to the treaty choice by the incumbent as a result of the election.

Furthermore, this numerical illustration contextualises the size of the office rent for the given examples. Essentially, R captures the relative importance of climate policy with respect to other policy aspects. Intuitively, if R is high, this means the party in office draws high levels of welfare from aspects unrelated to the climate policy choice. In order to capture this relative importance in a comparable way, the following ratio is introduced, where $W_i(R)$ corresponds to the welfare of the party in office:

$$\omega(R) = \frac{W_i(R=0)}{W_i(R=0) + R} \tag{47}$$

As an extreme example, $R = 0$ could thus be interpreted as climate policy making up 100% of an administration's welfare. As R increases, the value of ω decreases.

2.5.1 Green incumbent

The polarisation threshold values are given by $\bar{\phi} = 0.20$ and $\phi_G^M = 0.17$. With low polarisation (see Table 2), the threshold office rent that separates a consensus from a differentiation treaty roughly corresponds to a relative importance of climate policy of 25%. If incumbent draws less than 25% of their welfare from climate policy, they will opt for a differentiation treaty, which for the median voter, leads to a higher welfare loss.

Table 2: Low polarisation ($\phi = 0.15$), $\omega(\bar{R} = 1.23) \approx 0.25$

Treaty Type	Emission (e_i, e_j)		Reelection Pr.	ΔW_M^{loss}
Consensus	$\tilde{e}_1 = 0.891$	$\tilde{e}_1 = 0.891$	$p^C = 0.55$	0.008%
$\delta_i^* = 0.945 + \epsilon,\ R < \bar{R}$	$\tilde{e}_2 = 0.898$	$\tilde{e}_2 = 0.898$		
Differentiation	$\tilde{e}_1 = 0.891$	$\hat{e}_1 = 0.958$	$p^A = 0.5507$	0.107%
$\delta_i^* = 0.945 - \epsilon,\ R > \bar{R}$	$\tilde{e}_2 = 0.898$	$\hat{e}_2 = 0.95$		

For a higher level of polarisation (see Table 3), an assimilation treaty results since the challenger never ratifies. The incumbent optimally chooses a treaty that slightly reduces their reelection probability below the incumbency advantage: moving closer to the median voter's optimal treaty would prevent that, however, this is too costly in terms of weakening the treaty since that would mean foregoing some emission reductions by country 2.

Table 3: High polarisation ($\phi = 0.6$), $\omega(R = 0.5) \approx 0.33$

Treaty Type	Emission (e_i, e_j)		Reelection Pr.	ΔW_M^{loss}
Assimilation	$\tilde{e}_1 = 0.863$	$\hat{e}_1 = 0.98$	$p^A = 0.5497$	0.11%
$\delta_i^* = 0.938,\ R = 0.5$	$\tilde{e}_2 = 0.891$	$\hat{e}_2 = 0.95$		

2.5.2 Brown incumbent

Here, the polarisation threshold values are given by $\tilde{\phi} = 0.79$ and $\phi_B^M = 0.33$. For relatively low polarisation (see Table 4) the incumbent opts for a consensus over a differentiation treaty if the relative importance of climate policy is at least 40%. Choosing a differentiation treaty leads to a reelection probability above the incumbency advantage, because the challenger's non-cooperative emission choice is costly for the median voter, combined with comparably little emission reductions by country 2. Still, the median voter has a higher welfare loss with a differentiation treaty.

Table 4: Medium polarisation ($\phi = 0.6$), $\omega(\bar{R} = 1.12) \approx 0.40$

Treaty Type	Emission (e_i, e_j)		Reelection Pr.	ΔW_M^{loss}
Consensus	$\tilde{e}_1 = 0.966$	$\tilde{e}_1 = 0.966$	$p^C = 0.55$	0.66%
$\delta_i^* = 0.986 - \epsilon,\ R < \bar{R}$	$\tilde{e}_2 = 0.937$	$\tilde{e}_2 = 0.937$		
Differentiation	$\tilde{e}_1 = 0.966$	$\hat{e}_1 = 0.92$	$p^A = 0.5502$	0.735%
$\delta_i^* = 0.986 + \epsilon,\ R > \bar{R}$	$\tilde{e}_2 = 0.937$	$\hat{e}_2 = 0.95$		

High polarisation (see Table 5) means that no common ratification interval exists and the incumbent thus chooses between an insurance and a differentiation treaty. From the point of view of the median voter, the insurance treaty is slightly worse. Intuitively this is because indpendent of who wins the election, they get a bad outcome: the incumbent's non-cooperative emissions are very high, while due to the unambitious treaty, the treaty emissions are only slightly lower.

Table 5: High polarisation ($\phi = 0.8$), $\omega(\bar{R} = 0.58) \approx 0.31$

Treaty Type	Emission (e_i, e_j)		Reelection Pr.	ΔW_M^{loss}
Insurance	$\hat{e}_1 = 0.99$	$\tilde{e}_1 = 0.97$	$p^B = 0.5487$	0.775%
$\delta_i^* = 0.979,\ R < \bar{R}$	$\hat{e}_2 = 0.95$	$\tilde{e}_2 = 0.93$		
Differentiation	$\tilde{e}_1 = 0.971$	$\hat{e}_1 = 0.91$	$p^A = 0.5512$	0.748%
$\delta_i^* = 0.981,\ R > \bar{R}$	$\tilde{e}_2 = 0.932$	$\hat{e}_2 = 0.95$		

Generally speaking, higher levels of polarisation, both in case of a green and a brown incumbent, lead to higher welfare losses for the median voter. This can be attributed to the fact that high polarisa-

tion accentuates the incumbent's incentives to distort, combined with the fact that the median voter's preferences are by definition further removed from the ruling party's.

2.6 Extensions

While the basic model framework captures all of the main dynamics of the question at hand, it lends itself to a number of extensions. All of these have in common that they do not greatly affect the general insights presented, however, in some instances manage to add some nuance to the results.

2.6.1 Treaty Emissions as an Upper Bound

While we assume that instant renegotiation of a treaty is not possible, it could be argued that a government in power can always go beyond the promises made in an international treaty. Presumably, country 2 would not oppose to country 1 reducing emissions by more than what was agreed upon. While I would argue that this case is hardly seen in practice, illustrated by the lack of countries which overshoot their emission pledges, allowing for the elected government to go beyond treaty targets affects some outcomes of the model in an interesting fashion. In this extension, we will therefore interpret treaty emissions $\tilde{e}_1$ as an upper bound which the elected government can voluntarily undercut.

First, note that this change in assumption does not affect the equilibrium outcomes in the case of a brown challenger since the brown party's non-cooperative emission choice is never lower than treaty emissions:

$$\underbrace{1-\beta\theta_{j=B}}_{\hat{e}_{j=B}} < \underbrace{\delta_{i=G}(1-\beta\theta_{i=G})}_{\tilde{e}_{i=G}}, \qquad \text{since } \theta_{i=G} \geq \theta_{j=B} \text{ and } \delta_i \in [0,1].$$

Therefore, given that the brown challenger optimally wants to set higher emissions than the treaty emissions and due to the concavity of their welfare function, under a ratified treaty they cannot do better by choosing $e < \tilde{e}_{i=G}$.

Yet, it is possible for the green challenger to have lower non-cooperative emissions compared to treaty emissions. More precisely, this is the case whenever the treaty parameter is above a lower bound level δ^{LB}:

$$\begin{aligned} &\underbrace{1-\beta\theta_{j=G}}_{\hat{e}_{j=G}} \geq \underbrace{\delta_{i=B}(1-\beta\theta_{i=B})}_{\tilde{e}_{i=B}} \\ &\quad \Rightarrow \delta_i \geq \delta^{\text{LB}} \equiv \frac{1-\beta-\beta\phi}{1-\beta+\beta\phi}. \end{aligned} \tag{48}$$

Note that $\frac{d\delta^{\text{LB}}}{d\phi} < 0$, that is the more polarised the parties are, the lower this lower bound is.

Therefore, if the non-cooperative emission level is feasible ($\hat{e}_j \leq \tilde{e}_{1,i}$), which is the case whenever $\delta_{i=B} \geq \delta^{\text{LB}}$, the elected challenger will ratify the treaty and then choose $\hat{e}_j$. This makes sense intuitively: they cannot do better than to choose their individually optimal emission level, while at the same time getting the treaty benefit of country 2 reducing their emissions below their non-cooperative level, resulting in lower damage costs. Due to the assumption of a linear damage function, emission levels are dominant strategies and a lower than agreed emission level of country 1 does not affect the emission choice in country 2.

This now gives rise to two new cases *E* and *F*, which emerge depending on the level of polarisation as illustrated in Figure 15 and are distinguished by the order of $\underline{\delta}_i$ and δ^{LB}. If polarisation is low it holds that $\underline{\delta}_i < \delta^{\text{LB}}$. This means that there exists a common ratification interval, however, in which the green challenger chooses non-cooperative emissions after ratification. Case *F* thus indicates a range where, depending on who wins the election, *i* would ratify and set treaty emissions and *j* would ratify and choose $\hat{e}_j$.

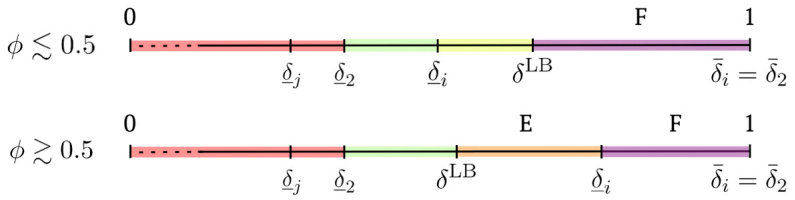


Figure 15: New cases *E* and *F* depending on polarisation levels

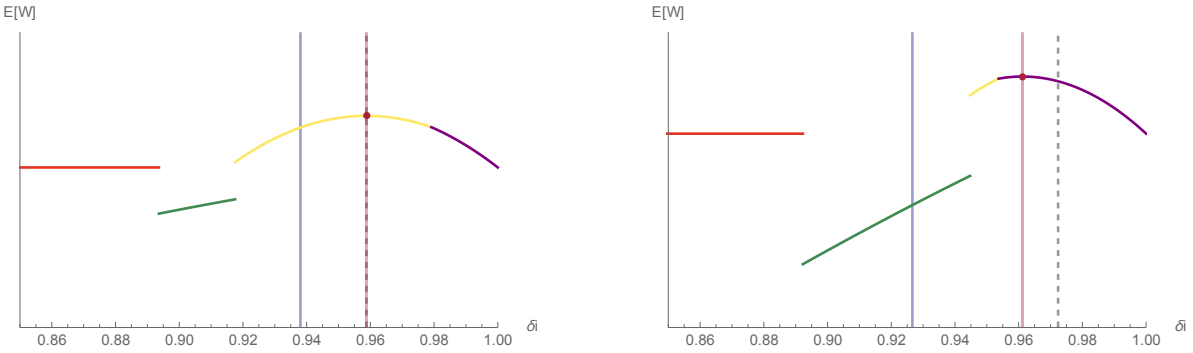
Increasing polarisation leads to the fact that $\underline{\delta}_i > \delta^{\text{LB}}$ and therefore there exists a range where only the green challenger ratifies (formerly case *B*) and then opts to choose non-cooperative emissions. In Case *E*, depending on who wins the election, *i* would thus not ratify and set non-cooperative emissions $\hat{e}_i$ and *j* would ratify and choose $\hat{e}_j$. Note that this outcome differs from case *D* in that country 2 will set treaty emissions.

Independent of polarisation levels, we find that classic differentiation treaties no longer exist. This is intuitive: the only way for a brown incumbent to differentiate from a green challenger in the basic model was to negotiate a treaty too weak for the challenger to ratify. Now, however, the challenger ratifies any treaty $\delta \in [\underline{\delta}_j, 1]$. The consensus and insurance treaty types still exist, albeit resulting from slightly different motives.

Low polarisation

With low polarisation levels, which here corresponds to $\phi \lesssim 0.5$, first-best and a variation of a consensus treaty are possible as seen in Figure 16. The former occurs when polarisation is sufficiently low. The latter differs from the original consensus treaty in the sense that we find ourselves in the range of case

F, where even though the challenger ratifies, they choose non-cooperative emissions. Still, country 2 engages in emission reductions as defined by the treaty. Note that even though $\hat{\delta}_i$ is available within case F, the incumbent optimally chooses a slightly more ambitious treaty. This is due to the fact that in case of election loss, the challenger will in any case choose non-cooperative emissions, however, country 2 will engage in more emission reductions if the treaty parameter is lower, which compensates for a slightly lower reelection probability.



(a) First-best treaty with $\phi = 0.2, R = 0.5$

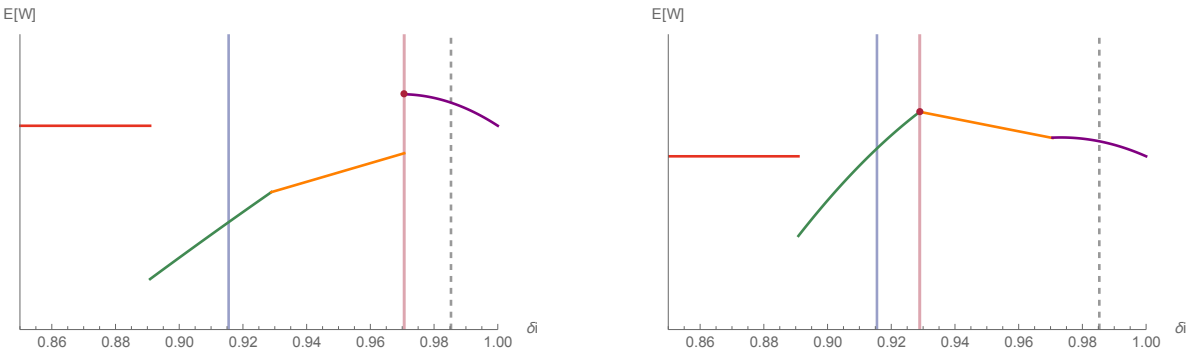
(b) "Consensus" treaty with $\phi = 0.45, R = 0.5$

Figure 16: Treaty outcomes with low polarisation

High polarisation

In the case of high polarisation, that is $\phi \gtrsim 0.5$, another difference to the main model materialises. Now the green challenger cannot be "locked" in with a weak treaty as before, since they can go beyond treaty emission reductions and thus the classic insurance motive is no longer available for the brown incumbent.

Analogously to the low polarisation case and for standard office rent levels, we observe a type of consensus treaty, in which both ratify but we are in area F, and where the incumbent opts for a treaty with $\delta < \hat{\delta}_i$ in order to achieve higher emission reductions by country 2. Now interestingly, as the office rent decreases and thus less importance is put on securing an election victory, a new treaty type emerges, as illustrated in Figure 17.



(a) "Consensus" treaty with $\phi = 0.7, R = 0.5$

(b) "Insurance" treaty with $\phi = 0.7, R = 0$

Figure 17: Treaty outcomes with high polarisation

The orange range indicates case E, where only the challenger ratifies the agreement but for any treaty parameter chooses non-cooperative emissions. In the basic model, the incumbent chose the upper limit of this range to lock in a cheap treaty because the challenger was bound to the treaty emissions, while now, they optimally propose the other end of the range δ^{LB}. At this point, treaty emissions exactly equal the green challenger's non-cooperative emissions. Intuitively, this is optimal because at any other point of the orange range, the challenger would also set non-cooperative emissions, while this is the point at which emission reductions by country 2 are maximised. Unchanged to the basic model, in case of an election win, the incumbent does not ratify the treaty. So conceptually, this treaty choice still resembles an insurance treaty in the sense that the incumbent aims at optimising the outcome in case of election loss.

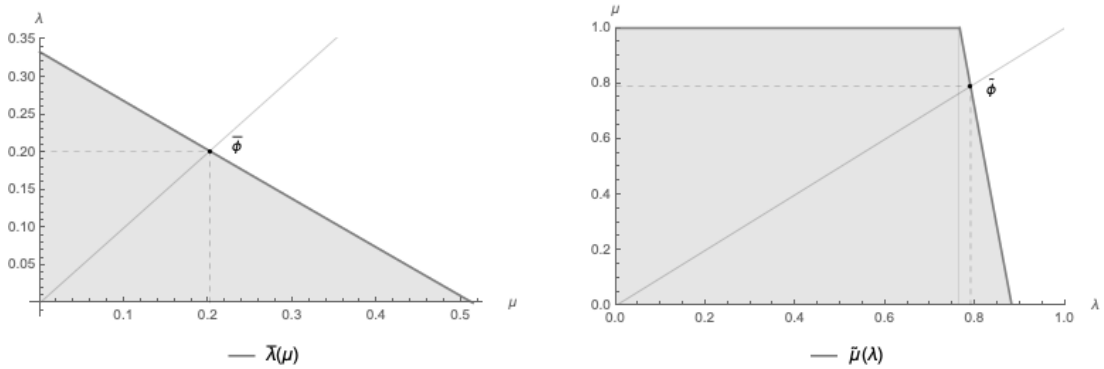
2.6.2 Preference Asymmetry

The basic model assumes the parties' preference parameters to be symmetric around the median voter. Allowing for preference asymmetry, that is for preference parameters that differ in terms of distance to the median voter, not only reproduces all of the presented results, but produces even more distorted outcomes. In particular, we will define preference parameters to be:

$$\theta_{1,G} = 1 + \mu, \qquad \theta_{1,B} = 1 - \lambda.$$

The degree of preference asymmetry is captured by the ratio of μ and λ, the symmetric cases being defined by $\mu = \lambda$. Emission choices still follow as described in Section 2.4.1. There are a few changes in the ratification stage from the results described in Section 2.4.2. Firstly, in the case of a green incumbent, the threshold value from Lemma 1 now becomes two-dimensional, as pictured in Figure 18a: Any combination of μ and λ within the shaded area ensures that the brown challenger's ratification interval exists, the boundary of which is the function $\bar{\lambda}(\mu)$. Similarly for the brown incumbent, the threshold

value to separate the two scenarios in Figure 7, as defined by Lemma 4, also becomes two-dimensional as illustrated in Figure 18b. The shaded area then indicates the parameter combinations for which an overlap in ratification intervals exists, the boundary being the function $\tilde{\mu}(\lambda)$.



(a) Threshold value analogy to $\bar{\phi}$

(b) Threshold value analogy to $\tilde{\phi}$

Figure 18: Two-dimensional analogy to threshold values $\bar{\phi}$ and $\tilde{\phi}$ for $\beta = 0.05$

For a more detailed and formal discussion of the changes in the ratification stage, refer to Appendix B.1. The election stage is qualitatively unaffected by the introduction of preference asymmetry. In the agreement stage, all of the outcomes presented in the symmetry case can also be replicated with asymmetry.

2.7 Conclusion

The interaction between political economy and international environmental cooperation has received comparably little attention in the literature so far. A deeper understanding of major political economy frictions is thus paramount for the creation of more successful treaties in the future. This model speaks to the importance of considering political polarisation in the context of domestic elections as a crucial element in international environmental policy.

To this aim, I investigate the role that domestic elections play for IEAs and to what extent they might be an explanatory factor for the modest success of international cooperation on climate change mitigation currently. Agents involved in international negotiations are often subject to domestic electoral concerns and therefore, policy decisions might affect their chances of reelection in upcoming elections, which consequently could affect their choice of optimal policy. Also, the presence of elections implies that the negotiation and the ratification decision might be made by two different governments.

I find that incumbent governments indeed have an incentive to distort their policy choices with the goal of affecting their chances in upcoming elections. These distortions can happen in three main ways:

the incumbent may negotiate (i) a *consensus treaty*, thereby making some concessions in order to secure ratification by the challenger and thus having a predictable environmental policy course, (ii) a *differentiation treaty* which exaggerates their preferences with the goal of offering two different policy choices to the median voter to improve reelection prospects, or lastly (iii) an *insurance treaty* that serves as an backstop against the challenger's non-cooperative policy choice. Which of the outcomes (i) – (iii) emerges, importantly hinges on the degree of polarisation between the two parties. Where treaty types (i) and (ii) correspond to results in Battaglini and Harstad (2020), and offer more of a microfoundation to their insights, agreements of type (iii) are novel.

The model setup includes a number of exogenous political economy factors: the office rent, the incumbency advantage and the density of the popularity shock all capture specific characteristics of a political system and therefore would allow for a discussion of the effects of domestic elections on international cooperation for a wide array of political landscapes. I find that the size of the office rent determines the weight that the incumbent puts on the maximisation of the reelection probability as opposed to achieving a favourable policy outcome. Similarly to Battaglini and Harstad (2020) this turns out to be the deciding factor between consensus and differentiation treaties. However, increased polarisation leads to the fact that consensus treaties are no longer available and that new scenarios emerge: a green incumbent chooses to assimilate towards the median voter, therefore weakening the treaty, while the brown incumbent opts for an insurance treaty, not committing to any emission reductions in case of reelection and handing a weak treaty to their challenger in case of election defeat.

Polarisation is generally found to be detrimental for the median voter's welfare and leads to a higher degree of policy distortion, in particular in the case of a brown incumbent. In some situations, however, even with high polarisation, the election pressure can lead to outcomes more preferred to the median voter than what would emerge in the absence of elections, as illustrated for consensus and assimilation treaties. In any case, this paper highlights the importance of paying attention to political polarisation within countries. I would argue that the relevance of this model expands beyond the specific application this model is set in: generally in the presence of cross-border public goods when there are separated governing bodies, the distorting incentives outlined will come into play. It can therefore be presumed that in the majority of cases, the underprovision of public goods is aggravated by local election pressure. Further research is therefore necessary in order to translate these findings into treaty mechanisms, which can, at least to some extent, overcome these inefficiencies.

Acknowledgements

I would like to thank Jean-Michel Benkert, Philipp Brunner, Nadia Ceschi, Simon Dietz, Igor Letina, Ralph Winkler and participants at the AURÖ Nachwuchsworkshop 2021 (Oldenburg), SAEE Workshop 2021 (Zurich), SURED 2022 (Ascona), EAERE 2022 (Rimini), and the Environmental Protection and Sustainability Forum 2022 (Graz) as well as seminar participants at the University of Bern and the London School of Economics for valuable comments. This research was supported by a SNSF Doc.Mobility Fellowship (P1BEP1_199981).

Appendix

A.2.1 Proofs

Proof of Lemma 1
(i) The challenger's threshold values exist if the term in the square root in (23), that is, $M_{j=B}$ is non-negative. Therefore:

$$\underbrace{\beta^2 \overbrace{(\beta-1)}^{<0}\overbrace{(\phi-1)}^{<0}}_{>0}\Big[1-5\phi+\beta(4\phi^2+5\phi-1)\Big] \geq 0$$

It thus suffices to consider the term in square brackets to determine the sign:

$$\begin{aligned} 1-5\phi+\beta(4\phi^2+5\phi-1) &\geq 0 \\ \phi &\leq \frac{5-5\beta+\sqrt{(\beta-1)(41\beta-25)}}{8\beta} \equiv \bar{\phi} \end{aligned}$$

(ii) Proof of (20):

$$\frac{d\bar{\phi}}{d\beta} = \frac{25-33\beta-5\sqrt{\beta-1)(41\beta-25)}}{8\beta^2\sqrt{\beta-1)(41\beta-25)}}$$

The term under the square root is non-negative since $\beta - 1 < 0$ and $41\beta - 25 < 0 \; \forall \beta \in [0, 0.15]$. The sign is thus determined by the numerator:

$$\begin{aligned} 25-33\beta-5\sqrt{\beta-1)(41\beta-25)} &\lesseqgtr 0 \\ (25-33\beta)^2 &\lesseqgtr 25(\beta-1)(41\beta-25) \\ 625+1089\beta^2-1650\beta &\lesseqgtr 625+1025\beta^2-1650\beta \\ 64\beta^2 &> 0 \end{aligned}$$

and therefore $\frac{d\bar{\phi}}{d\beta} > 0$.

□

Proof of Proposition 12
(i) Ratification interval for $i = G$:
Threshold values follow from (12), which is a concave function and has two roots. The upper ratification threshold is equal to 1 because it corresponds to the incumbent's non-cooperative emission choice and thus makes them equally well off as without a treaty.

To show when $\underline{\delta}^{i=G}$ is non-negative, consider the values of $\phi \in [0,1]$ that renders $\underline{\delta}^{i=G} = 0$:

$$\begin{aligned} &\underline{\delta}^{i=G} = 0 \\ \Rightarrow\ &\phi_0^{i=G} = \frac{2-\beta-\sqrt{3-4\beta+\beta^2}}{\beta} \end{aligned}$$

Knowing that the lower ratification threshold for the green incumbent decreases in ϕ (see Proposition 13) we can then show that ϕ_0 is only restrictive if marginal damages are sufficiently high:

$$\phi_0 \leq 1 \ \Rightarrow\ \beta \leq 0.1465$$

Consequently, as long as $\beta \leq 0.1465$, it holds that $\underline{\delta}^{i=G} \geq 0$.

(ii) Ratification interval for country 2:
Threshold values follow from (17), which is a concave function and has two roots.

The lower ratification threshold is always positive:

$$\begin{aligned} \underline{\delta}_2 = \frac{1+\beta[\beta(3+2\phi)-4]}{(\beta-1)^2} &> 0 \\ 1-4\beta &> 0 \qquad \forall \beta \in [0, 0.15] \end{aligned}$$

The upper ratification threshold is equal to 1 because it corresponds to their non-cooperative emission choice and thus makes them equally well off as without a treaty.

(iii) Ratification interval for $j = B$:
Threshold values follow from (12), which is a concave function and has two roots. We will now show that the lower ratification threshold is always positive. Since the denominator is quadratic, it suffices to look at the sign of the numerator:

$$1+\beta\left[\phi-3\right]-\beta^2\left[\phi^2+\phi-2\right]-\sqrt{M_{j=B}} \lesseqgtr 0$$

Note that ratification thresholds for the brown challenger only exist if $\phi \leq \bar{\phi}$, where $\bar{\phi}$ is maximised at $\beta = 0.15$ and $\bar{\phi}(0.15) \approx 0.206$ Also, $M_{j=B}$ is maximised at $\beta = 0.15$ and $\phi = 0$ and at those parameter values is $M_{j=B} \approx 0.02$, and consequently $\sqrt{M_{j=B}} \approx 0.14$. Therefore:

$$\underbrace{1+\beta\ \overbrace{\left[\phi-3\right]}^{\in[-3,-2.794]}\ -\beta^2\overbrace{\left[\phi^2+\phi-2\right]}^{\in[-2,-1.752]}}_{\in[0.589,0.6259]} > \underbrace{\sqrt{M_{j=B}}}_{\in[0,0.14]}$$

Since the two denominators do not overlap, the numerator and hence the lower ratification threshold of the brown incumbent is always positive.

Next, we will show that the upper ratification threshold is lower than 1. Restating it as follows:

$$\frac{A+\sqrt{M_{j=B}}}{B} \lesseqgtr 1,$$

and then reformulating:

$$\sqrt{M_{j=B}} \lesseqgtr B - A \equiv C \;\Leftrightarrow\; C^2 - M_{j=B} \lesseqgtr 0$$
$$4\beta^2\phi^2(\beta(1+\phi)-1)^2 > 0$$

Therefore, we have shown that $\frac{A+\sqrt{M_{j=B}}}{B} < 1$. □

Proof of Proposition 13

(i) For the green incumbent's thresholds:

$$\frac{d\underline{\delta}^{i=G}}{d\phi} = \frac{-2\beta \overbrace{(\beta-1)}^{<0}\overbrace{(1+\beta(1+\phi))}^{>0}}{\underbrace{(\beta(1+\phi)-1)^3}_{<0}} < 0 \tag{A.1}$$

The upper ratification threshold of the green incumbent is a constant and therefore not a function of ϕ.

(ii) For the brown challenger's thresholds:

- $\frac{d\underline{\delta}^{j=B}}{d\phi} > 0$

$$\frac{d\underline{\delta}^{j=B}}{d\phi} = \frac{\beta(\beta-1)\Big[\beta^2(\phi^2-15\phi-8)+\beta^3(2\phi^3-5\phi^2-10\phi+5)+3\sqrt{M_{j=B}}-}{(\beta(1+\phi)-1)^3\sqrt{M_{j=B}}}$$
$$\frac{\beta(-3+5\sqrt{M_{j=B}}+\phi(5+\sqrt{M_{j=B}}))\Big]}{(\beta(1+\phi)-1)^3\sqrt{M_{j=B}}} > 0 \tag{A.2}$$

We will show that this holds true. The denominator is negative. In the numerator, given that $\beta(\beta-1) < 0$, we will show that the term in square brackets is positive. Rewriting this term:

$$\sqrt{M_{j=B}}(\underbrace{3-5\beta-\beta\phi}_{A}) + \phi^3(\underbrace{2\beta^3}_{>0}) + \phi^2(\underbrace{\beta^2-5\beta^3}_{B}) + \phi(\underbrace{15\beta^2-10\beta^3}_{>0})$$
$$+\beta(\underbrace{3-5\phi+5\beta^2-8\beta}_{C})$$

where

$$A = 3 - 5\beta - \beta\phi > 0 \quad \text{for } \beta < \frac{3}{5-\phi} \in [0.58, 0.6], \text{ depending on } \phi \in [0, \tilde{\phi}].$$
$$B = \beta^2 - 5\beta^3 > 0 \quad \text{for } \beta < \frac{1}{5}$$
$$C = 3 - 5\phi + 5\beta^2 - 8\beta > 0 \quad \text{(as shown below)}$$

Note that the expression C is strictest at $\beta = 0.15$ and $\phi = \bar{\phi} \approx 0.206$, where $C(\beta = 0.15, \phi = 0.206) = 0.7925 > 0$. Consequently, since all term are positive, the expression in square brackets is also positive. Hence the numerator is negative, and combined with the negative denominator, it holds that $\frac{d\bar{\delta}^{j=B}}{d\phi} > 0$.

- $\frac{d\bar{\delta}^{j=B}}{d\phi} < 0$

$$\frac{d\bar{\delta}^{j=B}}{d\phi} = \frac{\beta(1-\beta)\Big[\beta^2(\phi^2+15\phi-8) + \beta^3(2\phi^3-5\phi^2-10\phi+5) - 3\sqrt{M_{j=B}} + \beta(3+5\sqrt{M_{j=B}} + \phi(-5+\sqrt{M_{j=B}}))\Big]}{(\beta(1+\phi)-1)^3\sqrt{M_{j=B}}} < 0 \tag{A.3}$$

Again, the denominator is negative, so that it remains to show that the numerator is positive. First, $\beta(1-\beta) > 0$. Then, rewriting the numerator:

$$A + [-3 + 5\beta + \beta\phi]\sqrt{M_{j=B}} \lesseqgtr 0$$

Note that the term in square brackets is negative. We want to show that the expression is > 0, now to get to the next line, we divide by the term in square brackets, which is why the sign flips and we now want to show that the expression is < 0.

$$M_{j=B} - \left(\frac{-A}{[-3+5\beta+\beta\phi]}\right)^2 \lesseqgtr 0$$

which simplifies to

$$\frac{1}{(\cdot)^2}\left(4\beta^2\phi \underbrace{(\beta(\phi-1)-1)^3}_{<0} \underbrace{[6 - \overbrace{5\phi}^{\max=1.03} - \overbrace{\beta(10-9\phi+\phi^2)}^{\max=1.5}]}_{>0}\right) < 0.$$

Therefore, the term in square brackets in the numerator is positive, making the whole expression negative.

□

Proof of Proposition 14

(i) To show that (24) is true, note that for $\phi = 0$ it holds that $\underline{\delta}^{i=G} = \underline{\delta}^{j=B}$ (because in this case $\theta_i = \theta_j$) and thus $\Delta\underline{\delta}^{i=G} = 0$. As stated in Proposition 13, the incumbent's lower threshold decreases in ϕ as shown in (A.1), while the challenger's lower threshold increases in ϕ as shown in (A.2). It thus follows that $\Delta\underline{\delta}^{i=G} \leq 0$.

(ii) By the same reasoning, the second relation of this Proposition, that is, (25), is true. At $\phi = 0$, it holds that $\bar{\delta}^{i=G} = \bar{\delta}^{j=B}$. Then, $\bar{\delta}^{i=G}$ is constant in ϕ, while $\bar{\delta}^{j=B}$ decreases with increasing ϕ, as shown in (A.3), meaning that $\Delta\bar{\delta}^{i=G} \geq 0$. □

Proof of Lemma 2

(i) To show that (26) holds true, first note that if $\phi = 0$, $\underline{\delta}^{i=G} = \underline{\delta}_2(\theta_G)$ because the two share the same environmental preference parameter. Then, as ϕ increases, the two values diverge as follows:

$$\frac{d\underline{\delta}_2}{d\phi} = \frac{2\beta^2}{(\beta-1)^2} \geq 0$$
$$\frac{d\underline{\delta}^{i=G}}{d\phi} \leq 0 \qquad \text{as shown in (A.1)}$$

Therefore, for any $\phi \in [0,1]$, $\underline{\delta}^{i=G} \leq \underline{\delta}_2(\theta_G)$.

(ii) To show that (27) holds true, a couple of steps are necessary. First, comparing the denominators of the two elements:

$$\bar{\delta}_2 = \frac{A}{(\beta-1)^2}$$
$$\bar{\delta}^{j=B} = \frac{B}{(\beta-1+\beta\phi)^2}$$

Since $(\beta-1)^2 \geq (\beta-1+\beta\phi)^2$, it holds that $\frac{1}{(\beta-1)^2} \leq \frac{1}{(\beta-1+\beta\phi)^2}$. Consequently, if $A \leq B$ is true, then $\bar{\delta}_2 \leq \bar{\delta}^{j=B}$ follows and (27) holds true. We therefore now show that $A \leq B$:

$$B - A \geq 0$$
$$\beta(\phi+1) - \beta^2\left(\phi^2+3\phi+1\right) - \sqrt{M_{j=B}} \geq 0$$
$$\phi\underbrace{(8-12\beta)}_{>6.2} + \phi^2\underbrace{(2\beta-4-12\beta^2)}_{>-4} + \phi^3\underbrace{(4\beta-10\beta^2)}_{>0} + \phi^4\underbrace{(-\beta^2)}_{>-0.0009} + 2\beta\underbrace{(1-\beta)}_{>0.85} \geq 0$$

Note that the positive terms in brackets are sufficient to cover the negative terms in brackets, even more so when taking into account the multiplications with ϕ. □

Proof of Proposition 15

(i) Ratification interval for $i = B$:

Threshold values follow from (12), which is a concave function and has two roots. The upper ratification threshold is equal to 1 because it corresponds to the incumbent's non-cooperative emission choice and thus makes them equally well off as without a treaty.

The lower ratification threshold is always positive, within the assumed parameter ranges:

$$\underline{\delta}^{i=B} = \frac{1+\beta(\phi-1)[4+\beta(\phi-3)]}{(1+\beta(\phi-1))^2} > 0$$

$$\underbrace{1+4\beta\phi+\beta^2\phi^2+3\beta^2}_{\in[1,1.56]} > \underbrace{3\beta^2\phi+4\beta+\beta^2\phi}_{\in[0,0.67]}$$

Given that the two intervals never overlap, this is always true.

(ii) Ratification interval for country 2:
Threshold values follow from (17), which is a concave function and has two roots. The upper ratification threshold is equal to 1 because it corresponds to country 2's non-cooperative emission choice and thus makes them equally well off as without a treaty.

The lower ratification threshold is always positive:

$$\underline{\delta}_2 = \frac{1+\beta[2\beta(1-\phi)+\beta-4]}{(\beta-1)^2} > 0$$

$$\underbrace{1+3\beta^2}_{\in[1,1.07]} > \underbrace{4\beta+2\beta^2\phi}_{\in[0,0.64]}$$

Given that the two intervals never overlap, this is always true.

(iii) Ratification interval for $j = G$:
Threshold values follow from (12), which is a concave function and has two roots. They exist if $M_{j=G}$ is non-negative:

$$M_{j=G} = \underbrace{\beta(1+\phi)(1-\beta)}_{>0}\underbrace{\left[1+5\phi+\beta(4\phi^2-5\phi-1)\right]}_{\mathcal{M}}$$

where

$$\mathcal{M} = \underbrace{1-\beta}_{>0}+5\phi\underbrace{(1-\beta)}_{>0}+4\beta\phi^4 > 0$$

and thus $M_{j=G} > 0$, hence the threshold values always exist.

To show when $\underline{\delta}^{j=G}$ is non-negative, consider the value of $\phi \in [0,1]$ that renders $\underline{\delta}^{j=G} = 0$:

$$\begin{aligned} &\underline{\delta}^{j=G} = 0 \\ \Rightarrow\ &\phi_0^{j=G} = \frac{2 - 2\beta - \sqrt{3 - 4\beta + \beta^2}}{\beta} \end{aligned}$$

Knowing that the lower ratification threshold for the green challenger decreases in ϕ (see Proposition 13), we can show that $\phi_0^{j=G}$ is only restrictive if marginal damages are sufficiently high:

$$\phi_0^{j=G} \leq 1 \ \Rightarrow\ \beta \leq 0.1465$$

Note that this is analogous to the condition in Proposition 12.

Also, the challenger's upper threshold is never above 1:

$$\bar{\delta}^{j=G} = \frac{1 - \beta\left[3 + \phi + \beta(\phi - 2)(1 + \phi)\right] + \sqrt{M_{j=G}}}{\left(1 + \beta(\phi - 1)\right)^2} \leq 1$$

Denoting the numerator as N and the denominator as D:

$$\begin{aligned} \frac{N}{D} \leq 1 \ &\Leftrightarrow\ N - D \leq 0 \\ -4\beta^2\phi^2\left(1 + \beta(\phi - 1)\right)^2 &\leq 0 \end{aligned}$$

which is always true. □

Proof of Proposition 16
(i) For the brown incumbent's thresholds:

$$\frac{d\underline{\delta}^{i=B}}{d\phi} = \frac{2\beta(\overbrace{\beta - 1}^{<0})(\overbrace{\beta(\phi - 1) - 1)}^{<0}}{\underbrace{\left(1 + \beta(\phi - 1)\right)^3}_{>0}} > 0 \tag{A.4}$$

The upper ratification threshold of the brown incumbent is not a function of ϕ.

(ii) For the green challenger's thresholds:

- $\frac{d\underline{\delta}^{j=G}}{d\phi} < 0$

$$\frac{d\underline{\delta}^{j=G}}{d\phi} = \frac{\beta(\beta-1)\Big[3\sqrt{M_{j=G}} + \beta(3+\beta(\phi^2-15\phi-8)+}{(1+\beta(\phi-1))^3\sqrt{M_{j=G}}}$$
$$\frac{\beta^2(5+10\phi-5\phi^2-2\phi^3) - 5\sqrt{M_{j=G}} + \phi(5+\sqrt{M_{j=G}})\Big]}{(1+\beta(\phi-1))^3\sqrt{M_{j=G}}} \tag{A.5}$$

Note that the denominator is positive, as seen in (A.4). In the numerator, given that $\beta(\beta-1) < 0$, we have to show that the term in square brackets is positive. Rewriting this term:

$$\sqrt{M_{j=G}}\overbrace{(3-5\beta+\beta\phi)}^{A} + \beta\overbrace{\left(3-8\beta+5\beta^2\right)}^{B} + \phi\overbrace{\left(5\beta-15\beta^2+10\beta^3\right)}^{C}$$
$$+\phi^2\underbrace{\left(\beta^2-5\beta^3\right)}_{D} + \phi^3\underbrace{\left(-2\beta^3\right)}_{E}$$

where

$$A = 3-5\beta+\beta\phi > 0 \quad \text{for } \beta < \frac{3}{5-\phi} \in [0.6, 0.71], \text{ depending on } \phi \in [0, \tilde{\phi}].$$
$$B = 3-8\beta+5\beta^2 > 0 \quad \text{for } \beta < 0.6$$
$$C = 5\beta-15\beta^2+10\beta^3 > 0 \quad \text{for } \beta < 0.5$$
$$D = \beta^2-5\beta^3 > 0 \quad \text{for } \beta < 0.2$$
$$E = -2\beta^3 < 0$$

All terms but E are positive. However, the negative impact of E is covered, e.g. by term C as follows:

$$\phi C - \phi^3 E = \phi(5\beta-15\beta^2+10\beta^3-(2\beta^3\phi^2)) = \phi(5\beta-15\beta^2+\beta^3(\overbrace{10-2\phi^2}^{>0})) > 0$$

Consequently, the term in square brackets is positive, making the numerator of (A.5) negative and hence the whole expression negative.

- $\frac{d\bar{\delta}^{j=G}}{d\phi} < 0$

$$\frac{d\bar{\delta}^{j=G}}{d\phi} = \frac{\beta(\beta-1)\Big[3\sqrt{M_{j=G}} + \beta(-3+\beta(8+15\phi-\phi^2)+}{(1+\beta(\phi-1))^3\sqrt{M_{j=G}}}$$
$$\frac{\beta^2(-5-10\phi+5\phi^2+2\phi^3) - 5\sqrt{M_{j=G}} + \phi(-5+\sqrt{M_{j=G}})\Big]}{(1+\beta(\phi-1))^3\sqrt{M_{j=G}}} \tag{A.6}$$

The denominator is positive, as seen in (A.4). Again, given that $\beta(\beta-1) < 0$, we have to show that the term in square brackets is positive. Rewriting this term:

$$A + [3 - 5\beta + \beta\phi]\sqrt{M_{j=G}} \lesseqgtr 0$$
$$M_{j=G} - \left(\frac{-A}{[3-5\beta+\beta\phi]}\right)^2 \lesseqgtr 0$$

which simplifies to

$$\frac{1}{(\cdot)^2}\left(4\beta^2\phi \underbrace{(1-\beta(\phi-1))^3}_{>0}\underbrace{[6+5\phi-\overbrace{\beta(10+9\phi+\phi^2)}^{\max=3}]}_{>0}\right) > 0.$$

Therefore, the term in square brackets is negative, making the whole expression negative.

□

Proof of Proposition 17
(i) To show that (31) is true, note that for $\phi = 0$ it holds that $\underline{\delta}^{i=B} = \underline{\delta}^{j=G}$ and thus $\Delta\underline{\delta}^{i=B} = 0$. As stated in Proposition 16, the incumbent's lower threshold increases in ϕ as shown in (A.4), while the challenger's lower threshold decreases in ϕ as shown in (A.5). It thus follows that $\Delta\underline{\delta}^{i=B} \geq 0$.

(ii) By the same reasoning, the second relation of this Proposition, that is, (32), is true. At $\phi = 0$, it holds that $\bar{\delta}^{i=B} = \bar{\delta}^{j=G}$. Then, $\bar{\delta}^{i=B}$ is constant in ϕ, while $\bar{\delta}^{j=G}$ decreases with increasing ϕ, as shown in (A.6), meaning that $\Delta\bar{\delta}^{i=B} \geq 0$. □

Proof of Lemma 3
(i) Proof of (33): First, if $\phi = 0$, $\underline{\delta}^{i=B} = \underline{\delta}_2(\theta_B)$ because the two share the same preferences. Then, as ϕ increases, the two values diverge as follows:

$$\frac{d\underline{\delta}_2}{d\phi} = \frac{-2\beta^2}{(\beta-1)^2} \leq 0$$
$$\frac{d\underline{\delta}^{i=B}}{d\phi} > 0 \qquad \text{as shown in (A.4)}$$

Therefore, for any $\phi \in [0, \tilde{\phi}]$, $\underline{\delta}^{i=B} \geq \underline{\delta}_2(\theta_B)$ and thus (33) holds.

(ii) To show that (34) holds true, a couple of steps are necessary. First, comparing the denominators of the two elements:

$$\begin{aligned} \bar{\delta}_2 &= \frac{A}{(\beta-1)^2} \\ \bar{\delta}^{j=G} &= \frac{B}{(\beta-1-\beta\phi)^2} \end{aligned}$$

Since $(\beta-1)^2 \leq (\beta-1-\beta\phi)^2$, it holds that $\frac{1}{(\beta-1)^2} \geq \frac{1}{(\beta-1+\beta\phi)^2}$. Consequently, if $A \geq B$ is true, then $\bar{\delta}_2 \geq \bar{\delta}^{j=B}$ follows and (34) holds true. We therefore now show that $A \geq B$:

$$\begin{aligned} A - B &\geq 0 \\ \beta(1-\phi) + \beta^2\left(3\phi + 3 - \phi^2 - 1\right) + \sqrt{M_{j=G}} &\geq 0 \\ \underbrace{(8 - 20\beta + 12\beta^2)}_{>0} + \phi\underbrace{(4 + 2\beta - 10\beta^2)}_{>0} + \phi^2\underbrace{(2\beta + 2\beta^2)}_{>0} + \phi^3\underbrace{(-\beta^2)}_{<0} &> 0 \end{aligned}$$

Note that $2\beta^2\phi^2 > \beta^2\phi^3$ and thus the whole expression is strictly positive. □

Proof of Lemma 4
The two scenarios are separated at the point where the ratification intervals touch, that is, where $\underline{\delta}_i = \bar{\delta}_j$. Solving this for ϕ yields:

$$\tilde{\phi} = \frac{\sqrt[3]{3\sqrt{3}\sqrt{(\beta-1)^3\beta^6(\beta(\beta(7\beta-15)+41)-25)}} - 2(\beta-1)^2\beta^3(4\beta-13) - \frac{(\beta-1)\beta^2(5\beta+1)}{\sqrt[3]{3\sqrt{3}\sqrt{(\beta-1)^3\beta^6(\beta(\beta(7\beta-15)+41)-25)}} - 2(\beta-1)^2\beta^3(4\beta-13)} + 4(\beta-1)\beta}{3\beta^2}$$

where $\frac{d\tilde{\phi}}{d\beta} < 0$, i.e., the higher environmental damages, the smaller the range for common ratification. □

Proof of Lemma 5
The reelection probability has to be interior, that is $p^l \in (0,1)$. Note that the reelection probability is increased versus the incumbency advantage,if $\Delta W_M^l > 0$ and vice versa for $\Delta W_M^l < 0$. In the first case, we thus have to ensure that:

$$\begin{aligned} \sigma\Delta W_M^l + z &< 1 \\ \sigma &< \frac{1-z}{\Delta W_M^l} \end{aligned}$$

while for $\Delta W_M^l < 0$ it has to hold that:

$$\sigma \Delta W_M^l + z > 0$$
$$\sigma < \frac{z}{-\Delta W_M^l} = \frac{z}{|\Delta W_M^l|}$$

Note that for high values of ΔW_M^l these conditions become harder to fulfil (since the upper limit is lower). Therefore, whichever case A–D leads to the highest value of difference in median voter welfare $|\Delta W_M^l|$ will be restrictive for the shock density σ.

□

Proof of Proposition 18
The welfare difference for the median voter (depending on the cases) is given by:

$$\begin{aligned}
\Delta W_M^A &= \delta_i - 0.5(1+\delta_i^2) + \beta\left[2 + \delta_i^2\theta_i - \delta_i(2+\theta_i)\right] + \\
&\quad \beta^2\left[\delta_i(1+\theta_i - 0.5\delta_i\theta_i^2) + \theta_j(0.5\theta_j - 1) - 1\right] \\
\Delta W_M^B &= 0.5(1-\delta_i)^2 + \beta\left[\delta_i(2+\theta_i - \delta_i + \beta^2\left[1+\theta_i + 0.5\theta_i^2(\delta_i^2-1) - \delta_i(1+\theta_i)\right]\theta_i) - 2\right] \\
\Delta W_M^C &= 0 \\
\Delta W_M^D &= \beta^2\left[\theta_i(1-0.5\theta_i) + \theta_j(0.5\theta_j - 1)\right]
\end{aligned}$$

□

Proof of Proposition 19
To prove that 1. is true, we will show that $W_i^A(\delta = \underline{\delta}_j) > W_i^A(\delta = \bar{\delta}_j)$, meaning that there exists a point in the lower area A that yields a higher expected welfare than the highest point in the upper area A. Rewriting:

$$\begin{aligned}
W_i^A(\underline{\delta}_j) - W_i^A(\bar{\delta}_j) &= p^A\left[B(\tilde{e}_1(\underline{\delta}_j)) - \theta_i D(\tilde{E}(\underline{\delta}_j)) + R\right] + (1-p^A)\left[\hat{W}_i(\theta_j)\right] - \\
&\quad \left(p^A\left[B(\tilde{e}_1(\bar{\delta}_j)) - \theta_i D(\tilde{E}(\bar{\delta}_j)) + R\right] + (1-p^A)\left[\hat{W}_i(\theta_j)\right]\right) \\
&= p^A \underbrace{\left[B(\tilde{e}_1(\underline{\delta}_j)) - B(\tilde{e}_1(\bar{\delta}_j)) + \theta_i D(\tilde{E}(\bar{\delta}_j)) - \theta_i D(\tilde{E}(\underline{\delta}_j))\right]}_{\mathcal{A}}
\end{aligned}$$

where

$$\mathcal{A} = \underbrace{\left(8\beta\phi - 4\beta^2\phi(2+\phi)\right)}_{(1)} \underbrace{\sqrt{M_{j=B}}}_{\geq 0 \text{ for } \phi \leq \bar{\phi}}$$

and $(1) > 0$ if $2 > \beta(2+\phi)$, which is strictest at $\beta = 0.15$ and $\phi = \bar{\phi}$, where it holds, making the whole expression positive.

□

Proof of Proposition 20

While we cannot solve explicitly for the optimal treaty parameter within area A ($\delta^*_{i,A}$), it is implicitly defined by the following expression:

$$\frac{dW_i^A}{d\delta_i} = \frac{dp^A}{d\delta_i}\left[\tilde{W}_i(\delta_i) - \hat{W}_i(\theta_j) + R\right] + \frac{d\tilde{W}_i}{d\delta_i}p^A = 0 \tag{A.7}$$

To see that $\delta^*_{i,A} \in (\hat{\delta}_i, \delta^*_M)$, we will evaluate (A.7) at $\hat{\delta}_i$ and δ^*_M and show that it is increasing in the former and decreasing in the latter, meaning that graphically, $\delta^*_{i,A}$ is located in between the two. Two prerequisites are necessary to show this.

First, note that the treaty parameter within case A that maximises the reelection probability coincides with the median voter's optimal treaty parameter:

$$\delta_A^{max} = \delta^*_M = \frac{1+\beta(\beta(1+\theta_i)-2-\theta_i)}{(1-\beta\theta_i)^2} \quad \text{where} \quad \frac{dp^A}{d\delta_i} = \begin{cases} > 0 & \text{for } \delta_i < \delta_A^{max} \\ < 0 & \text{for } \delta_i > \delta_A^{max} \end{cases}$$

Second, it holds that $\hat{\delta}_i < \delta^*_M$, because:

$$\delta^*_M - \hat{\delta}_i = \frac{2\beta\phi - \beta^2\phi(\phi+2)}{(\beta(\phi+1)-1)^2} > 0$$

because $2 > \beta(2+\phi)$ as seen in the proof of Proposition 19.

Now, note that $\hat{\delta}_i$ is defined by $\frac{d\tilde{W}_i}{d\delta_i} = 0$. The derivative (A.7) evaluated at $\hat{\delta}_i$ thus becomes:

$$\frac{dW_i^A}{d\delta_i}\Big|_{\delta=\hat{\delta}_i} = \frac{dp^A}{d\delta_i}\left[\tilde{W}_i(\delta_i) - \hat{W}_i(\theta_j) + R\right] > 0 \tag{A.8}$$

This is true because:

$$\tilde{W}_i(\delta_i) - \hat{W}_i(\theta_i) = \frac{\beta^2 \overbrace{\left(0.5 - \beta + 0.5\beta^2\right)}^{>0}(\phi+1)^2}{(\beta(\phi+1)-1)^2} > 0$$

and $\hat{W}_i(\theta_i) > \hat{W}_i(\theta_j)$. Also as shown, $\hat{\delta}_i < \delta^*_M$ implies that $\frac{dp^A}{d\delta_i} > 0$. The positive sign of (A.8) implies that $\hat{\delta}_i < \delta^*_{i,A}$.

In a next step, note that δ_M^* is defined by $\frac{dp^A}{d\delta_i} = 0$, as shown above. The derivative (A.7) evaluated at δ_M^* thus becomes:

$$\frac{dW_i^A}{d\delta_i}\Big|_{\delta=\delta_M^*} = \underbrace{\frac{d\tilde{W}_i}{d\delta_i}}_{<0} p^A < 0 \tag{A.9}$$

since $\frac{d\tilde{W}_i}{d\delta_i} = 0$ and $\hat{\delta}_i < \delta_M^*$. The negative sign of (A.9) thus implies that $\delta_M^* > \delta_{i,A}^*$. □

Proof of Proposition 21

(i) To prove that 1. is true, we will show that $W_i^B(\delta = \underline{\delta}_i) \leq W_i^C(\delta = \underline{\delta}_i)$, meaning that there always exists a point in area C which yields a weakly higher expected welfare for the incumbent and they thus never choose a treaty in area B. First, rewriting:

$$\begin{aligned} W_i^C - W_i^B &= B(\tilde{e}_1) - \theta_i D(\tilde{E}) + p^C R - \\ &\qquad \Big[p^B(B(\hat{e}_{1,i}) - \theta_i D(\hat{E}) + R) + (1-p^B)(B(\tilde{e}_1) - \theta_i D(\tilde{E}))\Big] \\ &= \left[B(\tilde{e}_1) - \theta_i D(\tilde{E})\right] p^B + R(p^C - p^B) - \left[B(\hat{e}_{1,i}) - \theta_i D(\hat{E})\right] p^B \\ &= p^B \left[\underbrace{\tilde{W}_i - \hat{W}_i}_{(1)}\right] + R \left[\underbrace{p^C - p^B}_{(2)}\right] \end{aligned}$$

We will now evaluate this difference at $\delta = \underline{\delta}_i$, where cases B and C meet. Now note that by definition of ratification threshold values, (1) is equal to zero: at $\underline{\delta}_i$, the incumbent is indifferent between ratifying or not, making the two values exactly equal.

Therefore, to show that $W_i^B(\delta = \underline{\delta}_i) \leq W_i^C(\delta = \underline{\delta}_i)$ holds, (2) has to be non-negative:

$$(p^C - p^B)\Big|_{\delta=\underline{\delta}_i} = \frac{4 - 4\phi + \beta\left(-2\phi^2 + 10\phi - 8\right) + \beta^2\left(2\phi^2 - 6\phi + 4\right)}{(1 + \beta(\phi - 1))^2}$$

Given that the denominator is quadratic, the numerator determines the sign of the expression.

$$\underbrace{4 - 4\phi + \beta\left(-2\phi^2 + 10\phi - 8\right)}_{(3)} + \beta^2 \underbrace{\left(2\phi^2 - 6\phi + 4\right)}_{\geq 0 \text{ for } \phi \leq 1}$$

Note that (3) is smallest at $\phi = 1$ and equals $2\beta - 2\beta^2$, meaning that (3) is non-negative for all values $\beta \geq 0$. Consequently, a treaty in area B is always weakly dominated by a treaty in area C.

(ii) Next, we show that 3. is true, that is for all values $\phi \leq \phi_B^M$, the optimal treaty parameter in the absence of an election is always in area C, making the first-best treaty always available. We will therefore

show that $\underline{\delta}_i \leq \hat{\delta}_i \leq \bar{\delta}_j$.

$$\hat{\delta}_i - \underline{\delta}_i = \frac{1}{(\cdot)^2}\left[\beta(1-\phi+\beta(\phi-1))\right] > 0 \qquad \text{for } \beta < 1 \text{ which is always true.}$$
$$\hat{\delta}_i - \bar{\delta}_j = \frac{1}{(\cdot)^2}\left[A - \sqrt{M_{j=G}}\right] \tag{A.10}$$

where the term in square brackets is negative whenever:

$$M_{j=G} - A^2 > 0$$
$$\beta^2(-4\phi^4+12\phi^3-15\phi^2+6\phi+1)+\beta(-12\phi^3+26\phi^2-12\phi-2)-11\phi^2+6\phi+1>0$$

Within the parameter range $\phi \in [0, \phi_B^M]$ this condition is strictest at $\phi = 0$ and $\beta = 0.15$, where it equals $0.725 > 0$. Therefore, for low polarisation levels, $\hat{\delta}_i$ is always within area C. □

Proof of Proposition 22
To prove that 1. is true, we will show that there always exists a point in area A that is weakly better than area D, ensuring that the incumbent never opts for a treaty that no party ratifies. First, by definition it holds that $W_i^A(\delta = 1) = W_i^D$. Now we will show that the function W_i^A takes its lowest point at $\delta = 1$, meaning it is decreasing in δ, and therefore there always exists a point in area A that is strictly better than a treaty parameter in area D.

$$\frac{dW_i^A}{d\delta}\Big|_{\delta=1} = \overbrace{(2\beta^3\sigma\phi^2 + R\beta\sigma)}^{>0} \underbrace{\left[\beta - 1 - \phi + \beta\phi - \beta\phi^2\right]}_{(1)} + z\beta \underbrace{\left[\phi - 1 + \beta(1-\phi)\right]}_{(2)} \tag{A.11}$$

where

$$(1) = \overbrace{\beta(1+\phi)}^{\leq 0.3} - 1 - \phi - \beta\phi^2 < 0$$
$$(2) = \phi(1-\beta) + \beta - 1 \leq 0 \text{ for } \phi \leq 0.$$

Consequently, (A.11) is negative. □

A.2.2 Extensions

Preference Asymmetry

Here we provide the formal background of the extension with preference asymmetry and show how the main Lemmas and Propositions are affected.

Green incumbent

Lemma 1b (Existence of Ratification Interval for Brown Challenger)

The challenger's ratification thresholds exist if:

$$\lambda \leq \bar{\lambda} = \frac{\beta - 1 + 2\mu(1 - \beta - \beta\mu)}{3(\beta - 1) + 2\beta\mu} \tag{B.1}$$

$$\mu \leq \bar{\mu} = \frac{1 - \beta - \sqrt{1 - 4\beta + 3\beta^2}}{2\beta} \tag{B.2}$$

and where it holds that:

$$\frac{d\bar{\lambda}}{d\mu} < 0, \quad \frac{d\bar{\mu}}{d\beta} > 0. \tag{B.3}$$

Proof of Lemma 1b

The challenger's threshold values exist if $M_{j=B}$ is non-negative. Therefore:

$$\underbrace{\beta^2 \overbrace{(\beta - 1)}^{<0} \overbrace{(\lambda - 1)}^{<0}}_{>0} [1 - 2\mu - 3\lambda + \beta(2\mu(1 + \mu) - 1 + \lambda(3 + 2\mu))] \geq 0$$

It thus suffices to consider the term in square brackets to determine the sign:

$$\begin{aligned} \lambda\,(3(\beta - 1) + 2\beta\mu) &\geq \beta - 1 + 2\mu\,(1 - \beta - \beta\mu) \\ \lambda &\leq \frac{\beta - 1 + 2\mu(1 - \beta - \beta\mu)}{3(\beta - 1) + 2\beta\mu} \equiv \bar{\lambda} \end{aligned}$$

Note that the sign switches between the two lines because the denominator is negative for the parameter range considered. This can be shown by assuming the parameters that maximise this expression, i.e. $\beta = 0.15$ and $\mu = 1$ and then $3(\beta - 1) + 2\beta\mu = -2.25 < 0$.

Note that at $\bar{\lambda}(\bar{\mu}) = 0$, meaning that $\mu \leq \bar{\mu}$ ensures that $\bar{\lambda}$ is non-negative.

Proposition 1b (Stage 3: Ratification Intervals with $i = G$)

In the case of a green incumbent, the incumbent's and country 2's ratification thresholds are given as

follows:

$$[\underline{\delta}^{i=G}, \bar{\delta}^{i=G}] = \left[\max\left\{0, \frac{1+\beta(1+\mu)\left[\beta(3+\mu)-4\right]}{(\beta+\beta\mu-1)^2}\right\}, 1\right] \tag{B.4}$$

$$[\underline{\delta}_2(\theta_G), \bar{\delta}_2(\theta_G)] = \left[\frac{1+\beta[2\beta(1+\mu)+\beta-4]}{(\beta-1)^2}, 1\right] \tag{B.5}$$

The challenger's ratification thresholds exist when $\lambda \leq \bar{\lambda}$ and $\mu \leq \bar{\mu}$. In that case, they are given by:

$$[\underline{\delta}^{j=B}, \bar{\delta}^{j=B}] = \left[\frac{1-\beta\left[3-2\lambda+\mu+\beta(\lambda-1)(2+\mu)\right]-\sqrt{M_{j=B}}}{(\beta+\beta\mu-1)^2}, \right.$$
$$\left. \min\left\{\frac{1-\beta\left[3-2\lambda+\mu+\beta(\lambda-1)(2+\mu)\right]+\sqrt{M_{j=B}}}{(\beta+\beta\mu-1)^2}, 1\right\}\right] \tag{B.6}$$

where $M_{j=B} = \beta^2(\beta-1)(\lambda-1)\left[1-2\mu-3\lambda+\beta(2\mu(1+\mu)-1+\lambda(3+2\mu)\right]$.

Proposition 2b (Stage 3: Comparative Statics with $i = G$)

The following conditions hold for the equilibrium ratification intervals under the condition that thresholds exist and that they are within the interval $[0, 1]$:

$$\frac{d\underline{\delta}^{i=G}}{d\mu} < 0, \qquad \frac{d\bar{\delta}^{i=G}}{d\mu} = 0, \qquad \frac{d\underline{\delta}^{i=G}}{d\lambda} = 0, \qquad \frac{d\bar{\delta}^{i=G}}{d\lambda} = 0$$

$$\frac{d\underline{\delta}^{j=B}}{d\mu} > 0, \qquad \frac{d\bar{\delta}^{j=B}}{d\mu} < 0, \qquad \frac{d\underline{\delta}^{j=B}}{d\lambda} > 0, \qquad \frac{d\bar{\delta}^{j=B}}{d\lambda} < 0$$

Propositions 14 & Lemma 2 hold analogously to the case of symmetry and as a consequence, cases as illustrated in Figure 6 follow. However, now the two scenarios ares distinguished by whether μ and λ are below threshold values as defined in Lemma B.1.

Brown incumbent

Proposition 4b (Stage 3: Ratification Intervals with $i = B$)

In the case of a brown incumbent, ratification thresholds are given as follows:

$$[\underline{\delta}^{i=B}, \bar{\delta}^{i=B}] = \left[\frac{1+\beta(\lambda-1)\left[4+\beta(\lambda-3)\right]}{\left(1+\beta(\lambda-1)\right)^2}, 1\right] \tag{B.7}$$

$$[\underline{\delta}_2(\theta_B), \bar{\delta}_2(\theta_B)] = \left[\frac{1+\beta(2\beta(1-\lambda)+\beta-4)}{(\beta-1)^2}, 1\right] \tag{B.8}$$

The challenger's ratification thresholds always exist and are given by:

$$[\underline{\delta}^{j=G}, \bar{\delta}^{j=G}] = \left[\max\left\{0, \frac{1+\beta\left[\lambda-3-2\mu-\beta(2-\lambda)(1+\mu)\right]-\sqrt{M_{j=G}}}{\left((1+\beta(\lambda-1)\right)^2}\right\}, \tag{B.9}$$

$$\frac{1+\beta\left[\lambda-3-2\mu-\beta(2-\lambda)(1+\mu)\right]+\sqrt{M_{j=G}}}{\left((1+\beta(\lambda-1)\right)^2}\right] \tag{B.10}$$

with $M_{j=G} = \beta^2(1-\beta)(1+\mu)\left(1+3\mu+2\lambda+\beta\left[2\lambda(\mu+\lambda-1)-3\mu-1\right]\right)$.

Proposition 5b (Stage 3: Comparative Statics with $i = B$)

The following conditions hold for the equilibrium ratification intervals under the condition that thresholds exist and that they are within the interval $[0, 1]$:

$$\frac{d\underline{\delta}^{i=B}}{d\lambda} > 0, \qquad \frac{d\bar{\delta}^{i=B}}{d\lambda} = 0, \qquad \frac{d\underline{\delta}^{i=B}}{d\mu} = 0, \qquad \frac{d\bar{\delta}^{i=B}}{d\mu} = 0$$

$$\frac{d\underline{\delta}^{j=G}}{d\lambda} \lessgtr 0, \qquad \frac{d\bar{\delta}^{j=G}}{d\lambda} < 0, \qquad \frac{d\underline{\delta}^{j=G}}{d\mu} < 0, \qquad \frac{d\bar{\delta}^{j=G}}{d\mu} < 0$$

Lemma 4b (Ordering of Countries' Ratification Intervals with $i = B$)

The two scenarios are separated at the point where the ratification intervals touch, i.e. where $\underline{\delta}_i = \bar{\delta}_j$. Solving this for μ yields the following threshold:

$$\tilde{\mu} = \frac{8-9\lambda-\beta(\lambda-1)(6\lambda-16+\beta\left[8+\lambda(\lambda-5)\right])}{(1+\beta(\lambda-1))^2} \tag{B.11}$$

where $\frac{d\tilde{\mu}}{d\lambda} < 0$. The threshold value is non-negative as long as $\lambda < \lambda_0$, where $\lambda_0(\beta)$ is most restrictive at $\beta = 0.15$ and takes a value of $\lambda_0(0.15) \approx 0.85$.

Proposition 17 and Lemma 3 hold analogously to the case of symmetry and consequently, cases as illustrated by Figure 7 follow.

References

Aldy, J. E., S. Barrett, and R. N. Stavins (2003). Thirteen plus one: a comparison of global climate policy architectures. *Climate policy 3*(4), 373–397.

Alesina, A. and G. Tabellini (1990). A positive theory of fiscal deficits and government debt. *The Review of Economic Studies 57*(3), 403–414.

Barrett, S. (1994). Self-enforcing international environmental agreements. *Oxford economic papers*, 878–894.

Barrett, S. (2002). Consensus treaties. *Journal of Institutional and Theoretical Economics*, 529–547.

Battaglini, M. and B. Harstad (2020). The political economy of weak treaties. *Journal of Political Economy 128*(2), 544–590.

Besley, T. and S. Coate (1998). Sources of inefficiency in a representative democracy: a dynamic analysis. *American Economic Review*, 139–156.

Buisseret, P. and D. Bernhardt (2018). Reelection and renegotiation: International agreements in the shadow of the polls. *American Political Science Review 112*(4), 1016–1035.

Carraro, C. and D. Siniscalco (1993). Strategies for the international protection of the environment. *Journal of Public Economics 52*, 309–28.

Cazals, A. and A. Sauquet (2015). How do elections affect international cooperation? evidence from environmental treaty participation. *Public Choice 162*(3-4), 263–285.

Finus, M. (2008). Game theoretic research on the design of international environmental agreements: insights, critical remarks, and future challenges. *International Review of environmental and resource economics 2*(1), 29–67.

Finus, M. and S. Maus (2008). Modesty may pay. *Journal of Public Economic Theory 10*, 801–26.

Gelman, A. and G. King (1990). Estimating incumbency advantage without bias. *American Journal of Political Science*, 1142–1164.

Hagen, A., J.-C. Altamirano-Cabrera, and H.-P. Weikard (2021). National political pressure groups and the stability of international environmental agreements. *International Environmental Agreements: Politics, Law and Economics 21*(3), 405–425.

Hänsel, M. C., M. A. Drupp, D. J. Johansson, F. Nesje, C. Azar, M. C. Freeman, B. Groom, and T. Sterner (2020). Climate economics support for the UN climate targets. *Nature Climate Change 10*(8), 781–789.

Harstad, B. (2022). Pledge-and-review bargaining: From Kyoto to Paris. *The Economic Journal*.

Hoel, M. (1992). International environment conventions: The case of uniform reductions of emissions. *Environmental and Resource Economics 2*, 141–59.

Köke, S. and A. Lange (2017). Negotiating environmental agreements under ratification constraints. *Journal of Environmental Economics and Management 83*, 90–106.

Kolstad, C. D. and M. Toman (2005). The economics of climate policy. *Handbook of environmental economics 3*, 1561–1618.

Levitt, S. D. and C. D. Wolfram (1997). Decomposing the sources of incumbency advantage in the us house. *Legislative Studies Quarterly*, 45–60.

List, J. A. and D. M. Sturm (2006). How elections matter: Theory and evidence from environmental policy. *The Quarterly Journal of Economics 121*(4), 1249–1281.

Marchiori, C., S. Dietz, and A. Tavoni (2017). Domestic politics and the formation of international environmental agreements. *Journal of Environmental Economics and Management 81*, 115–131.

Melnick, J. and A. Smith (2022). International negotiations in the shadow of elections. *Journal of Conflict Resolution*, 00220027221139433.

Persson, T. and G. Tabellini (1992). The politics of 1992: Fiscal policy and european integration. *The review of economic studies 59*(4), 689–701.

Persson, T. and G. Tabellini (1994). Representative democracy and capital taxation. *Journal of Public Economics 55*(1), 53–70.

Putnam, R. D. (1988). Diplomacy and domestic politics: the logic of two-level games. *International organization*, 427–460.

Robinson, J. A. and R. Torvik (2005). White elephants. *Journal of public economics 89*(2-3), 197–210.

Spycher, S. and R. Winkler (2022). Strategic delegation in the formation of modest international environmental agreements. *European Economic Review*, 103963.

Swiss Re (2021). The economics of climate change: no action not an option. *Swiss Re Management Institute 30*, 50.

UNEP (2022). Emissions gap report 2022: The closing window – climate crisis calls for rapid transformation of societies. *https://www.unep.org/emissions-gap-report-2022*.

Wagner, U. J. (2001). The design of stable international environmental agreements: economic theory and political economy. *Journal of Economic Surveys 15*, 377–411.

Chapter 3

Meet Me at the Threshold – Asymmetric Preferences in a Threshold Public Goods Game

Abstract: In this experiment, we analyse a threshold public goods game in which players have varying benefits from public goods provision, motivated by the existence of large heterogeneities between countries in international environmental cooperation. We find that provision is most frequent when players are symmetric. While increasing the degree of asymmetry does not significantly hamper provision success, contributions become more volatile the more heterogeneous players are. Analysing how players share contribution costs, we see that the extent of asymmetry is not salient, leading to relatively constant burden-sharing across treatments despite varied levels of inequity, often leading to allocations that diverge from both our benchmarks of efficiency or fairness.

3.1 Introduction

There exist many decision situations in which groups of people come together with the aim of realising a joint project, which only materialises if enough effort or contributions towards the common goal accrue. Crowdfunding is an example in which an organiser finances a project by setting a funding goal, and then implements it if enough funds accumulate. Often times in charity fundraising, the charity endeavour is only put into effect if there are sufficient donations. Both of these examples have in common that they can be captured by the structure of a *threshold public goods game* (TPGG). Players in this game contribute towards reaching a threshold, and the public good is only provided if this threshold is met. However, the public good can then be enjoyed by all players independently of their contribution level. Essentially it is thus a game of group effort, in which the members of the group might have varying preferences for the public good or different contribution costs, but can only ensure provision if they work together. Particularly when heterogeneities exist across players, it is not always trivial how the contribution burden should be split among group members.

Another prominent example of a high stakes group effort game is the mitigation of anthropogenic climate change, which is a collective action problem that requires international coordination and cooperation, a task impeded by the fact that no supranational entity can enforce a fair and efficient outcome. Unsurprisingly, achieving a consensus that is impactful has proven to be extremely difficult over the past decades. While it is scientifically undisputed that decisive and immediate action is needed to achieve the the widely pronounced policy goal of net-zero emissions by mid-century, current mitigation efforts, as pledged in the Paris Agreement, are not ambitious enough to achieve such a trajectory (UNEP 2022; IPCC 2022). Due to the fact that mitigation of climate change is impeded by the public goods property of greenhouse gas emission reductions and therefore plagued by freeriding incentives, the challenging nature of international cooperation on environmental policy is not surprising and we observe a constant underprovision of emission reductions. TPGGs resemble the real-world collective action problem of climate change mitigation, in which reaching a common target requires individual sacrifice whereas benefits only emerge if others contribute as well (Milinski et al. 2008), and might therefore provide valuable insights into how the global community could succeed in limiting the increase of surface temperatures to below $1.5^{\circ}C$ above pre-industrial levels in due time.

Many scholars have, by means of experimental analyses of public goods games, attempted to shed light on the relative importance of a variety of factors influencing the success of cooperation. One important aspect is the existence of large differences in the world community, for example with regards to wealth or exposure and vulnerability to climate change. Thus, equity considerations are playing an important role in achieving an impactful consensus in international environmental policy (Lange et al. 2010; Klinsky et al. 2017). One potential way of addressing such concerns is by analysing the behaviour of heterogeneous players in TPGGs. This paper thus discusses an experimental analysis of public goods provision, motivated by the difficulties of the international community in coordinating actions to mitigate climate change stemming from asymmetry between involved agents. In this spe-

cific setup, I analyse the effect of different preferences for the public good on whether the public good is provided and how contributions toward the threshold are split between players depending on the degree of asymmetry between them. In the one-shot two-player game, a continuum of Nash equilibria emerge from a theoretical point of view. The experiment thus offers insights into whether and on which equilibrium allocations players implicitly coordinate and how this is affected by asymmetric preferences and varying social value of the public good. Specifically I test how public goods preferences affect (i) the frequency of provision, (ii) the equilibrium selection, and (iii) the resulting allocation once the potentially focal equal split equilibrium is removed from the equilibrium set. I find that asymmetric players provide the public good less frequently than symmetric players, where the degree of asymmetry plays a less important role than hypothesised. Burden-sharing occurs in a way that can be considered unjust, following both efficiency and fairness considerations, observing too high contributions by "poorer" players and too low contributions by "richer" players. This implies that there might be a biased perception of heterogeneities between agents in collective action problems, leading to far from optimal burden-sharing in public goods provision.

TPGGs have been studied for a long time, both theoretically and experimentally. In the standard setting of the game, players simultaneously choose their individual contribution level. Public goods provision occurs, if a certain contribution threshold is reached. From a theoretical point of view, Bagnoli and Lipman (1989) discuss the "one-streetlight-problem", in which a group of neighbours wants to set up a streetlight and collect funds to do so. They essentially show that a "provision point mechanism" such as a TPGG achieves to alleviate the free-rider problem of public goods provision to some extent. In another early contribution, Palfrey and Rosenthal (1984) give the example of collecting money for a new office coffee pot, discussing the effect of differing refunding rules if not sufficient funds accrue.

In experimental setups, the specific rules of the game are very diverse and therefore only allow for limited comparison. To what extent contributions are multiplied in case of provision and whether contributions are wasted or (partially) refunded in case of non-provision, therefore varies depending on the specific research question and context (for an overview of early experiments see Croson and Marks 2000). Also, group sizes and whether games are repeated or one-shot heavily vary and influence findings. Here I will focus on contributions related to heterogeneity among players both in a general context as well as related to climate change. Within-group heterogeneity has been found to influence player's willingness to contribute to collective goods (Rapoport and Suleiman 1993; McGinty and Milam 2013; Fischbacher et al. 2014; Gavrilets 2015), where players can be unequal with respect to their wealth (endowments), their contribution costs, as well as to how non-provision affects their welfare. The degree to which heterogeneity affects successful provision strongly interacts with the specific rules of the game and experiment design, where both positive and negative effects are possible.

A very active and recent literature on TPGGs is specifically set in an environmental context, mirroring the fact that costly mitigation pledges in international environmental agreements are public goods and that there are uncertainties concerning the amount of contributions necessary to make a treaty sufficiently stringent. While even with a known threshold coordination among players proves to be

difficult, threshold uncertainty has received a lot of attention and is found to be detrimental to cooperation, since players tend to inaction when contributions are more risky (McBride 2010; Barrett and Dannenberg 2012; Dannenberg et al. 2015). Another complicating factor for coordination is players having differentiated stakes in the game, which is is where we put our main focus.

Evidence on how heterogeneity across agents affects the likelihood of public goods provision is also mixed in the environmental literature. Tavoni et al. (2011) show that inequality in individual endowments hampers successful provision. While in their experiment the threshold level is certain, damages from not reaching it materialise only with a probability of 50%. Additionally, they allow for communication in the form of pledges, which is shown to promote coordination. However, they find that the "poor" subjects are not willing to compensate for inaction of the "rich" and thus the paper underlines the importance for early leadership by the wealthy. Waichman et al. (2021), to some extent contrasting the findings from Tavoni et al. (2011), show that heterogeneity does not necessarily lead to lower success rates in public goods provision. In their study, they differentiate between two types of heterogeneity: wealth (endowment) and expected loss heterogeneity and find that in the latter specification, the success rate in meeting the threshold is higher than under symmetry, concluding that heterogeneity might not only be an obstacle. Burton-Chellew et al. (2013) also investigate a double heterogeneity setting, however, suggest that if there exists heterogeneity in wealth, groups are less likely to reach the threshold if the poorer participant are more heavily affected from climate damages. The experiment of Feige et al. (2018) analyses the effect of a non-binding voting procedure, where players are heterogeneous with respect to marginal contribution costs and play a repeated TPGG with an uncertain threshold in groups of four. Similarly to my setup, multiple equilibria exist which differ in the way contributions are split between players. They find that the predominant burden-share is that of equal contribution costs, which in their experiment also coincides with equal payoffs.

This paper adds a number of novel insights to the literature. To our knowledge we are the first to analyse a TPGG with heterogeneous public good preferences in combination with quadratic contribution costs. Also, analysing the effect of player heterogeneity in one-shot games without communication allows for a detailed discussion of the isolated role of asymmetric preferences as well as of varying social value of the public good. The convex cost specification allows for the investigation of the trade-off experimental subjects face between efficiency and equity. Even though convex contribution costs pose a conceptual challenge to participants, this degree of complexity is within reason due to the provision of a payoff calculator, which supports subjects in their understanding of the payoff structure. Finally, from a methodological point of view, the application of a sequential matching procedure is novel and renders it possible to generate a high number of observations compared to if players were playing in groups.

3.2 Theoretical Background

I consider a threshold public goods game with a threshold level T that is known with certainty. The game is played one-shot among two players called 1 and 2, in the following indexed by i, j, where $i \in \{1, 2\}$ and $j \neq i$. Players may have different benefits from public goods provision, captured by the preference parameter θ_i, but they have identical quadratic contribution costs. In case contributions are not sufficient to reach the threshold, players lose a fraction $q \in [0, 1]$ of their investment. The specific payoff functions are given as follows:

$$U_i(\theta_i, x_i) = \begin{cases} bT\theta_i - \frac{1}{2}cx_i^2 & \text{if } x_i + x_j \geq T \\ q(-\frac{1}{2}cx_i^2) & \text{otherwise} \end{cases} \qquad i, j = 1, 2,\ j \neq i,\ c > 0. \tag{1}$$

If the contribution of the other player is above the threshold T, it is optimal for player i to contribute zero. In case of $x_j < T$, there exists a cut-off value $\underline{x}_j$ which determines the minimum contribution of the opponent player such that it's a best response of agent i to contribute a positive amount. Above the cut-off value, the best response is to contribute such as to just reach the threshold. Below this cut-off value, the best response is not to contribute. In summary:

$$x_i = \begin{cases} 0 & \text{if } x_j \geq T \\ T - x_j & \text{if } \underline{x}_j \leq x_j < T \\ 0 & \text{if } x_j < \underline{x}_j \end{cases} \qquad i, j = 1, 2\ j \neq i, \tag{2}$$

where the cut-off value is type-dependent and is implicitly defined as the contribution level at which a player has a payoff of zero:

$$\begin{aligned} &U_i(T - x_j|\text{'provision'}) = U_i(0|\text{'no provision'}) \\ &\quad bT\theta_i - \frac{1}{2}c(T - x_j)^2 = 0, \qquad i, j = 1, 2\ j \neq i. \end{aligned} \tag{3}$$

This leads to the following cutoff-values and maximum contribution levels of both players:

$$\underline{x}_i(\theta_j) = T - \sqrt{\frac{2b}{c}T\theta_j}, \qquad \bar{x}_i(\theta_i) = \sqrt{\frac{2b}{c}T\theta_i} \qquad i, j = 1, 2\ j \neq i. \tag{4}$$

Hence, there exists a continuum of equilibria, all of which just reach the threshold and, in addition, the no contribution equilibrium $(x_i, x_j) = (0, 0)$. The continuum is characterised as follows:

$$\mathcal{C} = \left\{ (x_i, x_j) \mid x_i \in \{\underline{x}_i, \bar{x}_i\}, x_j \in \{\underline{x}_j, \bar{x}_j\},\ x_i + x_j = T \right\}. \tag{5}$$

Figure 19 shows the equilibria of the game with symmetric players, that is $\theta_i = \theta_j$, indicated by the fact that cut-off values are symmetric as well. The grey line illustrates the continuum (5).

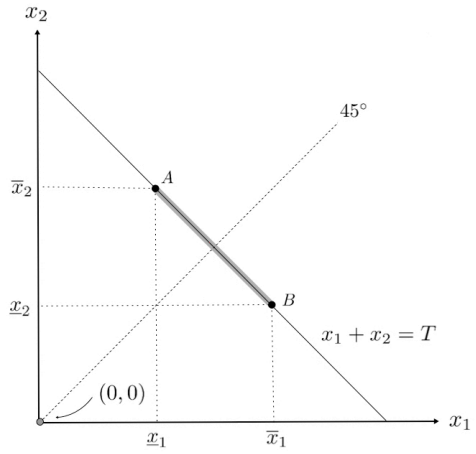


Figure 19: Illustration of the equilibrium interval and the no-contribution equilibrium

The equilibrium interval narrows as the valuation of the public good decreases (lower values of θ), as stated by (6) and (7).

$$\text{Point A:} \qquad \frac{d\underline{x}_j(\theta_i)}{d\theta_i} < 0\,, \qquad \frac{d\bar{x}_i(\theta_i)}{d\theta_i} > 0. \tag{6}$$

$$\text{Point B:} \qquad \frac{d\underline{x}_i(\theta_j)}{d\theta_j} < 0\,, \qquad \frac{d\bar{x}_j(\theta_j)}{d\theta_j} > 0. \tag{7}$$

3.2.1 Focal points

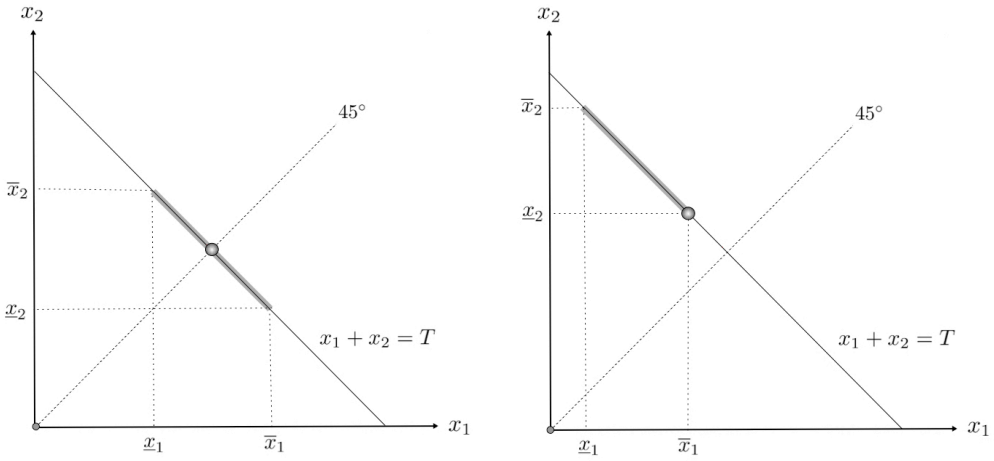
While all the points on the interval are equilibria, some of them can be interpreted as focal. In this section, I will highlight two such points, which are the efficient allocation that maximises joint payoff as well as an allocation which can be considered as *fair*, since it equally splits the gains from cooperation.

Efficiency

I define efficiency as the allocation that maximises joint payoff of players 1 and 2. Due to the fact that both players have the same quadratic contribution costs, the efficient allocation is given by the *equal split allocation*, that is $x_i = x_j = \frac{T}{2}$, as long as the benefits of provision outweigh the costs, that is when $bT(\theta_i + \theta_j) > 2c(\frac{T}{2})^2$. This can be seen by maximising joint payoffs given as follows:

$$\max_{x_i, x_j} \sum_{i=1}^{2} U_i(\theta_i, x_i) = (\theta_i + \theta_j)bT - \frac{1}{2}c(x_1^2 + x_2^2) \qquad \text{s.t. } x_1 + x_2 \geq T,\ (x_1, x_2) \in \mathcal{C} \tag{8}$$

Maximising joint payoff is a standard cost minimisation problem subject to the constraint that the threshold is reached. First, it is efficient to reach the threshold with precision, since any contribution beyond is wasteful. Second, the cost function $C = \frac{1}{2}c(x_1^2 + x_2^2)$ reaches a minimum at $x_1 = x_2$ and thus it is efficient for both to contribute half of the threshold. However, throughout this paper, we will refer to the efficient Nash equilibrium as the efficient allocation. Consequently, the equal split allocation is efficient, as long as it is in the continuum $\mathcal{C}$ of Nash equilibrium.



(a) Symmetry **(b)** (Strong) Asymmetry

Figure 20: Efficient allocation illustrated by $\circ$

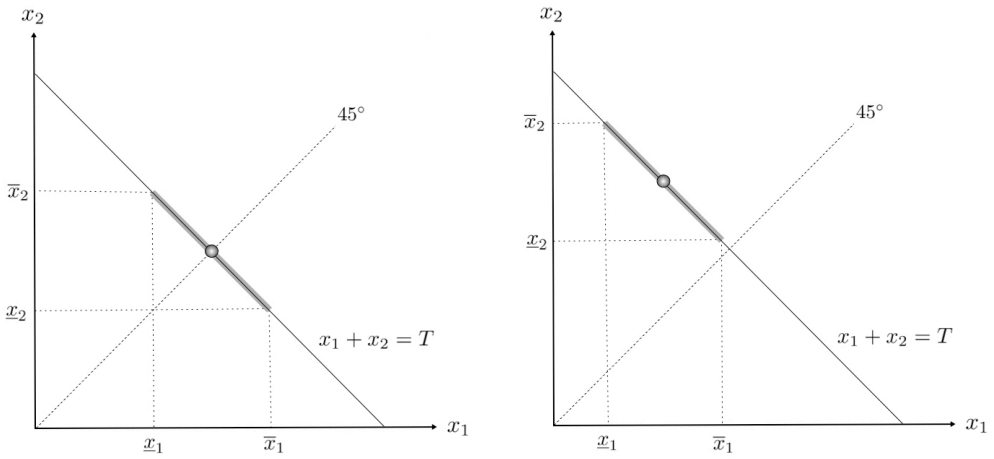
Figure 20 illustrates that this corresponds to the intersection of the 45° line with the threshold line. This allocation can be reached in equilibrium, as long as the equilibrium interval covers this point. In case of sufficiently heterogeneous preferences for the public good, this allocation, however, is not a Nash equilibrium. In this case, the most cost-efficient Nash equilibrium is that at the boundary of the equilibrium closest to the 45° line. Lastly, if public goods preferences are sufficiently small, the only and thus efficient equilibrium is the no-contribution allocation. The efficient allocation, depending on the degree of preference asymmetry can thus be summarised by:

$$\left(x_1^{eff}, x_2^{eff}\right) = \begin{cases} \left(\frac{T}{2}, \frac{T}{2}\right) & \text{if } \frac{T}{2} \in (\underline{x}_i, \bar{x}_i),\ i = 1,2 \\ (\bar{x}_1, T - \bar{x}_1) & \text{if } \frac{T}{2} \notin (\underline{x}_i, \bar{x}_i),\ \mathcal{C} \neq \varnothing,\ i = 1,2 \text{ and } \theta_2 > \theta_1 \\ (T - \bar{x}_2, \bar{x}_2) & \text{if } \frac{T}{2} \notin (\underline{x}_i, \bar{x}_i),\ \mathcal{C} \neq \varnothing,\ i = 1,2 \text{ and } \theta_1 > \theta_2 \\ (0,0) & \text{otherwise.} \end{cases} \tag{9}$$

Fairness

A possible conception of fairness is that players equally share the surplus generated from reaching the threshold, which can be a second focal point. This is given by the center of the equilibrium interval, given that it exists, as illustrated in Figure 21 and given by (10). Otherwise, the fair allocation equals the no-contribution allocation.

$$\left(x_1^{fair}, x_2^{fair}\right) = \begin{cases} \left(\frac{\bar{x}_1+\underline{x}_1}{2}, \frac{\bar{x}_2+\underline{x}_2}{2}\right) & \text{if } \underline{x}_i \geq 0 \qquad i = 1,2 \\ (0,0) & \text{otherwise} \end{cases} \tag{10}$$



(a) Symmetry **(b)** (Strong) Asymmetry

Figure 21: Fair allocation illustrated by ∘

Considering Figures 20 and 21 we can see that while in the symmetric case the two points overlap, as soon as there is preference asymmetry between the two players the values diverge. Both focal points will be used as benchmarks in the analysis of the experimental data in Section 3.4.

Finally, note that there potentially exist different conceptions of fairness among players, which would imply that this focal point is not the same for different players. One additional example could be that of equal cost sharing, which in this specification would correspond to equal contributions, coinciding with our definition of efficiency, as long as equal contribution is a Nash equilibrium.

3.3 Experimental Design

In the experiment, two subjects interact as players 1 and 2, choosing contribution levels x_1 and x_2 with the goal of reaching a threshold level T. Each pair plays 25 rounds, where each round differs with respect to assigned preference parameters for the public good θ_1 and θ_2. The experiment can be

interpreted as 25 rounds of one-shot games: players do not receive any feedback on their co-player's contribution and hence whether provision was successful until after all 25 rounds have been played. By randomising the order of rounds, learning and sequence effects are controlled for (see Section 3.4.1).

In the experimental setup, both T and b are normalised to 1 without loss of generality, while the cost parameter c is set to 10. Further, we set $q = 0.1$, meaning that players lose 10% of their contributions if the threshold is not reached. Contributions can be interpreted as percentage share of the investment. This simplifies the maximum contribution level according to (4) as follows:

$$\bar{x}_i = \sqrt{0.2\theta_i}, \qquad i = 1, 2. \tag{11}$$

Consequently:

$$\underline{x}_i = 1 - \sqrt{0.2\theta_j}, \qquad i, j = 1, 2,\ j \neq i. \tag{12}$$

The equilibrium interval (5) thus becomes:

$$\mathcal{C} = \left\{ (x_i, x_j) \;\middle|\; x_i \in \left\{ 1 - \sqrt{0.2\theta_j}, \sqrt{0.2\theta_i} \right\}, x_j \in \left\{ 1 - \sqrt{0.2\theta_i}, \sqrt{0.2\theta_j} \right\},\ x_i + x_j = T \right\}. \tag{13}$$

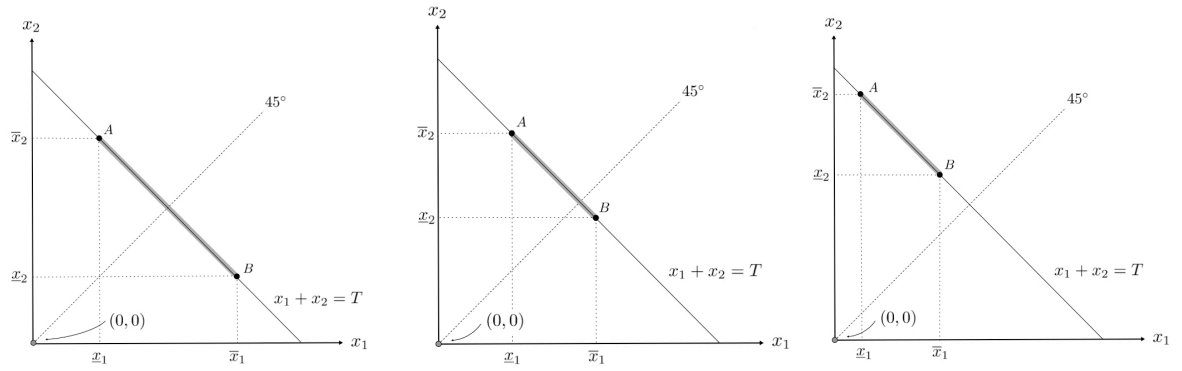
This interval reduces to a single point if $\bar{x}_1 = \bar{x}_2 = 0.5$, which holds for $\theta_1 = \theta_2 = 1.25$. Hence, if $\theta_i < 1.25$, the equal split allocation is outside of the equilibrium interval. Conversely, the full diagonal would be part of the equilibrium interval if $\bar{x}_1 = \bar{x}_2 = 1$, which is the case for $\theta_1 = \theta_2 = 5$.

The set of assigned preference parameters is given by $\theta_i \in \{0.75, 1.25, 2.5, 3.75, 4.25\}$, where all possible combinations among two subjects are played across the 25 rounds. This implies three different types of treatments in terms of preference asymmetry (illustrated in Figure 22), defined as follows:

- Symmetry (5 rounds): $\theta_1 = \theta_2$,
- Asymmetry (12 rounds): $\theta_1 \neq \theta_2$ and $\theta_i \neq 0.75$ for $i = 1, 2$,
- *Strong* asymmetry (8 rounds): $\theta_1 \neq \theta_2$ and $\theta_i = 0.75$ for either $i = 1, 2$.

Furthermore, this specification allows to disentangle the effects of preference heterogeneity and varying aggregate value of the public good. The latter relates to the fact that successful provision is expected to be easier if players value the public good more on aggregate. In order to isolate the effect of preference heterogeneity, five treatments keep the aggregate value of the public good constant at $\theta_1 + \theta_2 = 5$, and therefore constitute a special set of games:

- Constant social value: $\theta_1 + \theta_2 = 5$
  $(\theta_1, \theta_2) = (2.5, 2.5), (0.75, 4.25), (1.25, 3.75), (3.75, 1.25), (4.25, 0.75)$



(a) Symmetry **(b)** Asymmetry **(c)** Strong Asymmetry

Figure 22: Treatment types

Within this set, all three treatment types are represented. Note that for half of the remaining 20 treatments, the aggregate value of the public good is either higher or lower than 5. In Table 10 in the Appendix, a full overview over the 25 games is provided.

3.3.1 Hypotheses

The general aim of the experiment is to determine:

(i) How often the threshold is reached (success rate) and with which precision (deviation from threshold),

(ii) how contributions x_1 and x_2 compare (burden-sharing),

(iii) and how contributions compare to focal equilibrium points (deviation from fair/efficient).

I analyse how these insights are affected by the width and location of the equilibrium interval (5). Successful public good provision implies some implicit coordination on how to split contributions towards the threshold. This implicit coordination may be facilitated by focal points, where both efficiency and fairness, as defined in Section 3.2.1, may serve as such. As stated, under symmetry the two focal points coincide, while under asymmetry they diverge, conceivably hindering coordination. On top of that, in strongly asymmetric games, the equal split allocation is not available and thus no longer coincides with either of the two focal points, making coordination even more difficult. This leads to the following hypothesis:

Hypothesis 1 (Coordination and Preference Asymmetry)
Coordination is easier in symmetric games than in asymmetric games, making successful public good provision more frequent in symmetric games. Coordination is hardest and thus provision least frequent in strongly asymmetric games.

In addition, implicit coordination may be more difficult the larger the continuum $\mathcal{C}$ of Nash equilibria, as there are more options to profitably deviate from one of the focal points. However, at the same time a larger equilibrium interval also implies higher benefits (social value) of successful public goods provision, which should facilitate reaching the threshold. Therefore, we have two counteracting forces, which might harm or improve success rates, leading to the second hypothesis:

Hypothesis 2 (Coordination and Social Value)
The relation between increasing social value of the public good and the success rate of public good provision is non-monotonic.

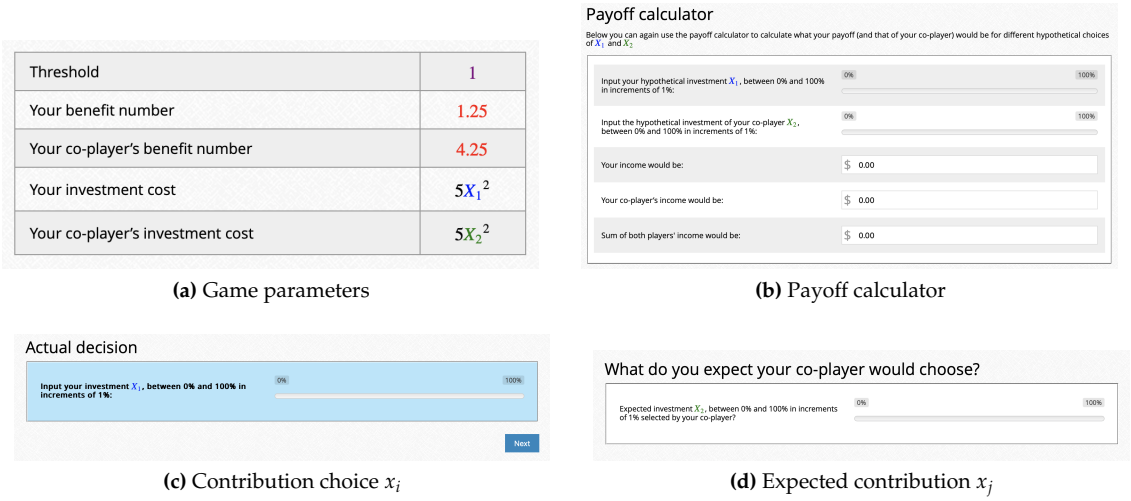
3.3.2 Implementation

The experiment was programmed by Expilab Research and consisted of four main phases. In the first phase, participants expressed consent and commitment to participate in the study and were then shown detailed instructions for the experiment and informed about compensation (fixed fee of 5£, bonus payment depending on average payoff across 25 rounds). In the next phase, participants had to complete three rounds of a tutorial aimed at testing their grasp of the provided payoff calculator. The third and main phase of the study consisted of the 25 rounds of the game. In each round, participants were given the parameters of the game (which changed across rounds) as well as a payoff calculator, as depicted in Figure 23. The payoff calculator provided two sliders for contribution levels x_1 and x_2 ranging from 0% to 100% in 1% increments. At the beginning of each round, the sliders were not set to any value in order to avoid framing effects (as pictured in Figure 23b). This payoff calculator allowed them to easily determine their own, their co-player's and joint payoff depending on contribution levels. For each round, players then had to choose an actual contribution level x_i and also indicate, what contribution x_j they expect their co-player to make[17].

After the subjects completed 25 rounds of the game, the final phase followed in the form of a questionnaire. Here, participants indicated demographic information and were asked about various aspects of the game. An overview of questionnaire replies is given in Section 3.4.1. Finally, participants received a completion code, which allowed them to retrieve compensation for participating in the study. After rounds 8 and 16 attention checks were included to ensure continuous attention of experimental subjects. The questions were completely unrelated to the experiment and concerned favourite fruits and cities. If a subject failed an attention check, the experiment was discontinued, which occurred one time.

As players played 25 rounds of a one-shot game without feedback after each round, it was not necessary for players to play the game simultaneously (as of calendar time) in pairs. In order to determine each player's payoff at the end of the 25 rounds, all players were matched with the player who at their respective start time had most recently completed the experiment. The contribution choices of this

[17] Due to technical issues, the expectation of x_j was not recorded throughout the experiment and can thus not be used in the data analysis.



(a) Game parameters

(b) Payoff calculator

(c) Contribution choice x_i

(d) Expected contribution x_j

Figure 23: Screenshots from main phase of experiment

matched co-player thus served as j values for player i. This, however, meant that some participants were chosen as the co-player of multiple other subjects, giving their contribution choices a higher weight in the total dataset, while some players' contribution choices never appear as player j. While this was the matching procedure relevant for the computation of the bonus payments, the data analysis will be based on a randomised matching in which each player is randomly determined to be one other participant's co-player.

3.4 Results

The results of the experiment will be reported in three parts. First, an overview of the subject pool will be provided with respect to demographics as well as detailing participants' responses in the post-experiment questionnaire. Also, a potential learning effect over the course of the experiment will be discussed. Second, a focus is put on symmetric treatments in order to address the interaction of interval width and the success rate and thus Hypothesis 1. Third, I will discuss constant social value games, asymmetric games as well as strongly asymmetric games separately in order to focus on the isolated effect of preference asymmetry from different angles, addressing Hypothesis 2.

3.4.1 Overview

The experiment was conducted on 20 January 2023 on the platform Prolific with a subject pool of 106 participants based in the United Kingdom.

Subject Pool

The subject panel is balanced in terms of gender and highly educated, with more than 60% of participants having completed higher education. Figure 24 details the exact distribution of gender and education. The mean age is 42 years, with the youngest participant being 20 and the oldest 71 years old. 89% of participants indicated English to be their first language.

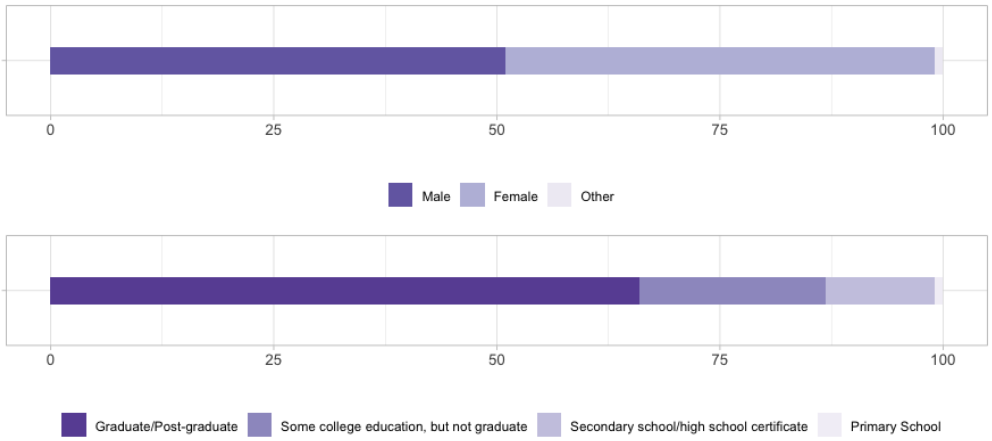


Figure 24: Gender and education distribution

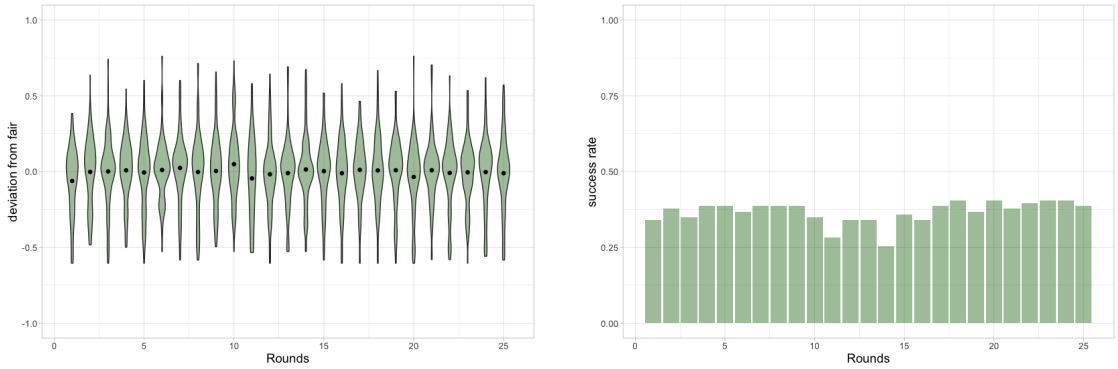
In the post-experiment questionnaire, participants on average indicated the level of difficulty to be slightly above average with a mean of 3.69 on a scale of 0 (extremely easy) to 6 (extremely difficult). When asked about risk attitudes, more than half of participants stated to be between risk neutral to risk prone, with the average being at 2.9 on a scale from 0 (extremely risk prone) to 6 (extremely risk averse). This might seem surprisingly high, however, it has to be kept in mind that these are self-reported values rather than deducted from a lottery. The subject pool quite evenly distributed on a scale from "very rarely" to "very often" when asked about whether they tend to trust people. Almost all participants indicated to have donated to charity before, with a frequency evenly ranging from "monthly" to "at most 5 times in my life". Finally, when asked about their skills at working with fractions, subjects averaged a score of 3.63 on a scale from 0 (not good at all) to 6 (extremely good).

Participants were also asked about their strategies and their perception of several aspects of the game. A majority of subjects (53%) indicate that the allocation they perceive as fair is the one in which players choose contributions such as to maximise joint payoff. Interestingly, this corresponds to what we define as the efficient focal point. However, as pointed out in Section 3.2.1, this could also be interpreted as "cost fairness". Roughly a quarter of subjects (23%) consider the allocation which ensures similar payoffs to both players as fair, which is also our definition of the fair focal point. Furthermore, 57% think that contributions should be such that both players have a positive payoff, if possible. When asked about picking the most difficult aspect of the game, the most frequent answers were "grasping the differences between rounds" (37%) and "guessing contribution X_2" (32%). The most important

rationale for contribution decisions for most players (49%) was "group efficiency", which was defined as maximising the joint income of players, followed by "monetary self-interest" (17%) and "avoiding risk" (14%), the latter being defined as avoiding being alone with a high contribution, potentially risking not reaching the threshold. A more detailed picture of the post-experiment survey can be found in the appendix.

Sequence Effect

The different variants of the game were played in a randomised order for each subject, which allows us to interpret each game as a one-shot game. However, it could theoretically be possible that subjects (i) experience a learning effect from playing the game and therefore their performance might systematically improve over the rounds, or that (ii) player's attention or motivation decreases, leading to deteriorating performances. This is not what we see in the data, illustrated by Figure 25.



(a) Deviation from fair

(b) Success rate

Figure 25: No sequence effect over rounds

Figure 25a shows the distribution of deviations from the fair contribution in each round of the game, a measure comparable across different treatments, with the black dots indicating means per round. Due to the fact that the order of games was randomised, in each round a random mix of games was played. If there was a sequence effect from playing the game for 25 rounds, we could expect to see a trend, which is not what is depicted. This visual impression is confirmed by one-way ANOVA tests, with the null hypotheses being that there is no difference between means across rounds. The null hypothesis is clearly not rejected ($p = 0.890$). The same exercise can be done with the deviation from the efficient contribution, where a one-way ANOVA test also clearly indicates that means are not significantly different ($p = 0.594$), see Figure 48 in Appendix C. A second variable of interest is the success rate, which is computed as the percentage share of games in which the threshold was successfully reached. If subjects experienced a sequence effect, they might have reached the threshold more or less frequently as rounds progressed. Figure 25b depicts average success rates per round, the difference between which

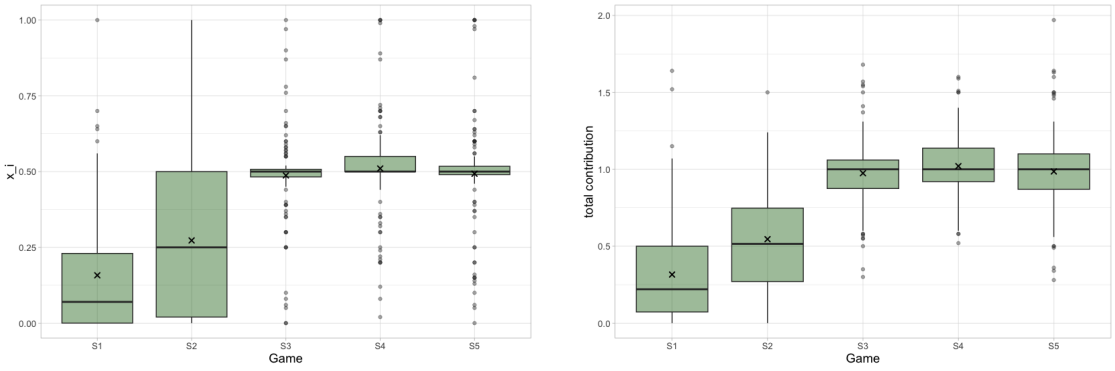
is not statistically significant (one-way ANOVA, $p = 0.304$). We therefore conclude that there is no significant sequence effect during the course of the study.

3.4.2 Symmetry

In this section, I will focus on the five symmetric games, which will be referred to as games S1–S5. Note that the preference parameters in S1 are so low that an efficient provision of the public good is not possible. In S2, only the equal split and the no contribution allocations ensure non-negative payoffs. For the remaining three games, it is efficient and fair to choose the equal split allocation. Table 6 provides summary statistics. We can see that for games S3–S5, the mean and medium contribution levels of player i are practicially indistinguishable, which is confirmed by pairwise two-sided Wilcoxon rank-sum (WRS) tests (S3 & S4 $p = 0.176$, S3 & S5 $p = 0.842$, S4 & S5 $p = 0.241$). Furthermore, the spread of contributions in games S1 and S2 is visibly higher, as seen in Figure 26. Note that the mean is indicated with a cross, where the bar corresponds to the median.

Table 6: Summary statistics for symmetric treatments

Game	Type	(θ_i, θ_j)	mean x_i	median x_i	sd x_i
S1	S	(0.75, 0.75)	0.1578	0.0700	0.2107
S2	S	(1.25, 1.25)	0.2728	0.2500	0.2301
S3	S	(2.5, 2.5)	0.4873	0.5000	0.1623
S4	S	(3.75, 3.75)	0.5102	0.5000	0.1701
S5	S	(4.25, 4.25)	0.4929	0.5000	0.1904



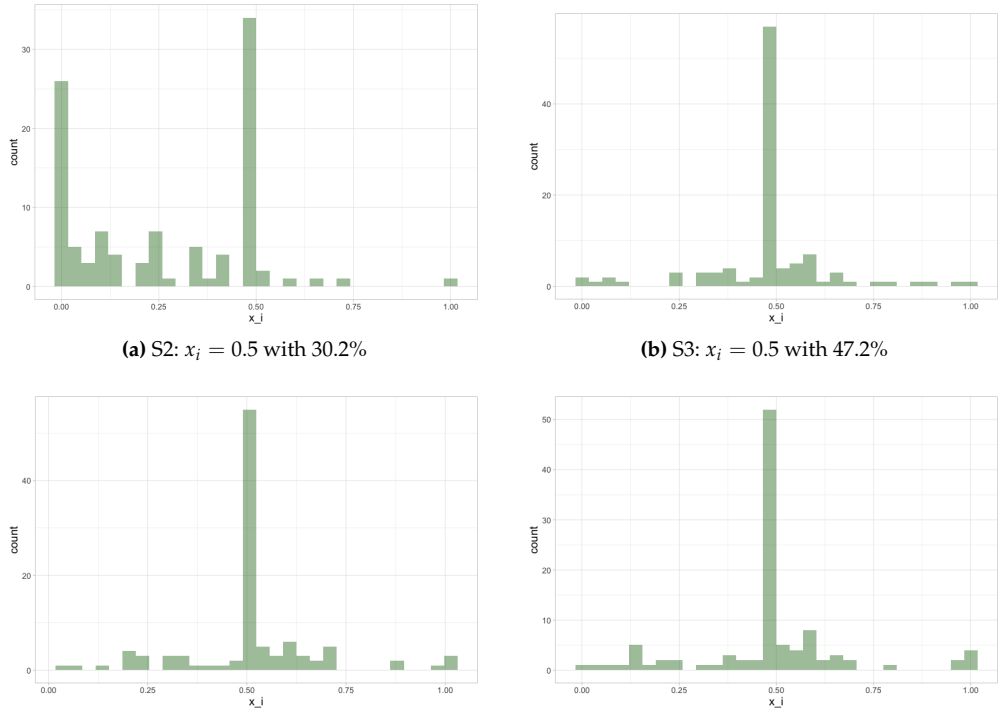
(a) Contributions x_i

(b) Total contributions $x_i + x_j$

Figure 26: Distribution of contributions in symmetric games

The equal split allocation in symmetric games corresponds to the fair and efficient focal point when available part of the equilibrium interval. In Game 1, only a contribution of zero ensures a non-negative

payoff, which is chosen most frequently ($x_i = 0$ with 38.7%). For the other four games, a contribution of 0.5 is on the equilibrium interval and as illustrated in Figure 27 is focal. Note that in S2, contributions of zero and the equal split are equivalent in terms of payoffs, that is, both equal to zero. Indeed, participants chose those two levels with similar frequency ($x_i = 0$ with 22.6%).



(a) S2: $x_i = 0.5$ with 30.2%

(b) S3: $x_i = 0.5$ with 47.2%

(c) S4: $x_i = 0.5$ with 48.1%

(d) S5: $x_i = 0.5$ with 40.6%

Figure 27: Contribution frequencies in the symmetric games

Figure 28 illustrates the success rates for the symmetric games. Remember that in S1 it is not possible to reach the threshold without incurring a negative joint payoff, where in S2 only the equal split allocation ensures a non-negative joint payoff, justifying the low success rates. We can see that the success rate increases up until S4, but then decreases for S5. The difference in the success rate, however, is not statistically significant between games S4 and S5 (one-sided WRS test, $p = 0.100$). Increasing stakes, that is higher social value, tend to increase provision success more than a narrower interval seems to facilitate coordination. Still, it is notable that the success rate does not further increase after S4, despite the fact that all focal points coincide, suggesting a potential non-linear relationship between the interval width and the success rate, as suggested in Hypothesis 2.

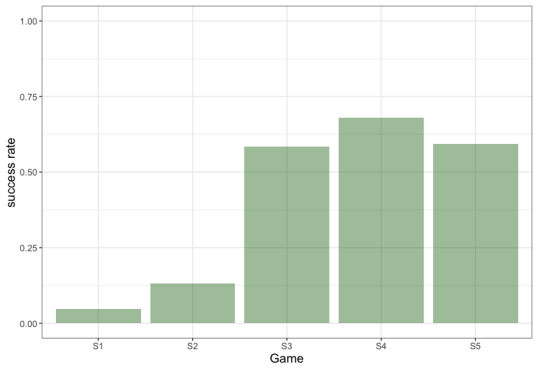
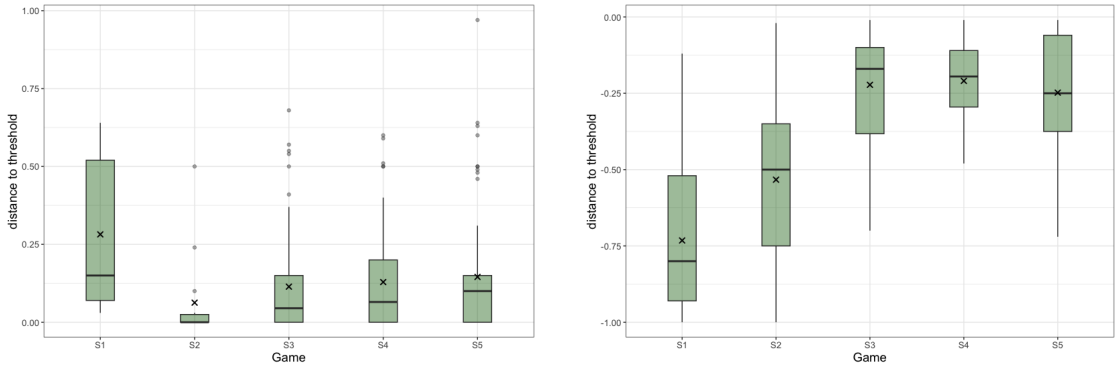


Figure 28: Success rate in symmetric games

Figure 29b shows the groups which failed to meet the threshold and depicts the distance distribution from total contributions to the threshold. The substantial distance in S1 and S2 is not surprising, but interestingly in S5 the threshold was missed by more on average than in the two other games (means S3–S5: -0.2223, -0.2094, -0.2479), even though in this game the social value of the public good was the highest.



(a) Distance to threshold if succeeded

(b) Distance to threshold if failed

Figure 29: Success and failure in symmetric games

Figure 29a analogously show the distribution of distances to the threshold for successful groups. Note that the surprisingly high value for S1 can be attributed to the fact that success in this game only occurred five times and thus is based on outlier values. The spread of S2 contributions is very low and most values close to zero, mirroring the fact that only the equal split yields non-negative payoffs. For S3–S5, the mean overshoot ranges from 0.114 to 0.145.

3.4.3 Effect of Asymmetry

This section highlights the effect of asymmetry from three different angles. First, keeping the social value of the public good constant, I investigate how different degrees of asymmetry affected success rates and burden-sharing among players. Second, focussing on the asymmetric treatment type, I discuss how increasing one player's stake in the game affects success and contributions. Finally, the same exercise will be conducted for strongly asymmetric games.

Constant Social Value Games

We define the *constant social value* (CSV) treatments as the five games in which the sum of θ values is equal to 5, and therefore the games in which the effect of asymmetry is isolated. Note that CSV3 is identical to game S3 discussed in the previous section. The summary statistics of the five games are given in Table 7 and contribution levels are illustrated in Figure 30. In all five games it is possible to reach the threshold with a positive joint payoff.[18]

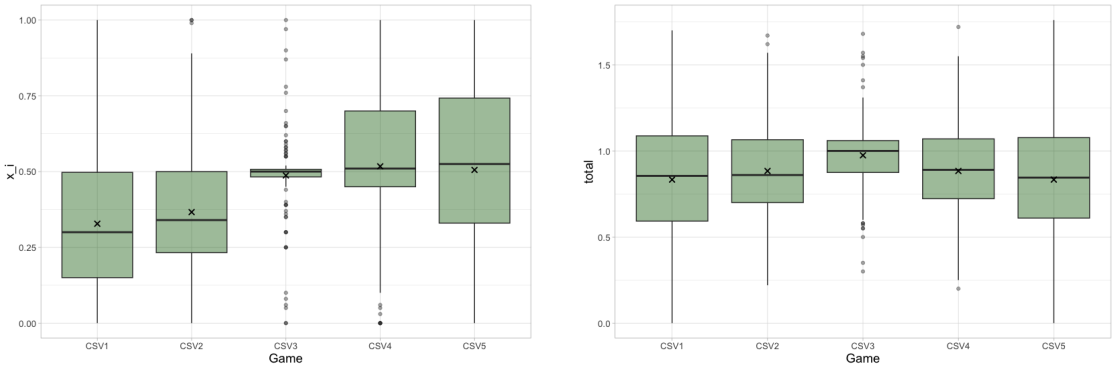
Table 7: Summary statistics for CSV treatments

Game	Type	(θ_i, θ_j)	mean x_i	median x_i	sd x_i
CSV1	SAS	(0.75, 4.25)	0.3279	0.3000	0.2597
CSV2	AS	(1.25, 3.75)	0.3663	0.3400	0.2277
CSV3	S	(2.5, 2.5)	0.4873	0.5000	0.1623
CSV4	AS	(3.75, 1.25)	0.5173	0.5100	0.2308
CSV5	SAS	(4.25, 0.75)	0.5055	0.5250	0.2833

Figure 30a essentially pictures two groups of games between which contribution levels clearly differ, that is games CSV1 and CSV2 in which asymmetry is against player i and the remaining games CSV3–CSV5. The mean contribution level also significantly varies within the first group (one-sided WRS test, $p = 0.043$) as well as between CSV4 and the two others (pairwise one-sided WRS test, CSV3 & CSV4 $p = 0.015$, CSV3 & CSV5 $p = 0.041$). The spread of contributions is lowest in the symmetric game.

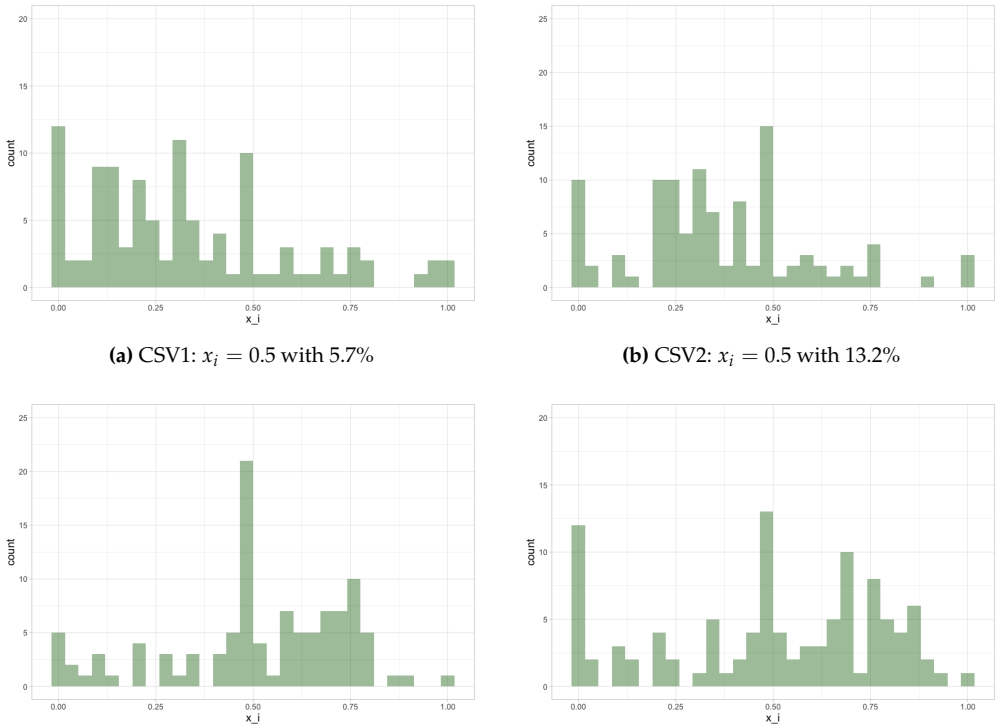
As seen in Figure 30b, on average, total contributions are highest in the symmetric game CSV3, with the difference being significantly different to all four games (pairwise one-sided WRS test, highest $p = 0.014$). The equal split contribution is less frequent than in AS and SAS games than in CSV3, as seen by comparing the frequency of $x_i = 0.5$ contributions between Figure 27b and 31. While, with the exception of CSV1, the equal split is still the contribution level with the highest count in asymmetric games, however, only with a maximum share of 16% compared to 47.2% in Game CSV3.

[18] Note that theoretically speaking, games CSV1/CSV5 and CSV2/CSV4 are *mirror* games. However, because subjects did not play in fixed pairs, results based on x_i and x_j are not exactly identical for mirror games. Example: Player a played CSV1 with their co-player b, but player b played CSV5 with *their* co-player c.



(a) Contributions x_i **(b)** Total contributions $x_i + x_j$

Figure 30: Distribution of contributions in CSV games.



(a) CSV1: $x_i = 0.5$ with 5.7% **(b)** CSV2: $x_i = 0.5$ with 13.2%

(c) CSV4: $x_i = 0.5$ with 16.0% **(d)** CSV5: $x_i = 0.5$ with 11.3%

Figure 31: Contribution frequencies in the asymmetric constant social value games

The difference in success rates between symmetric and asymmetric/strongly asymmetric games is statistically significant (two-sided WRS test, pair-wise between all games, max. $p = 0.0093$). The difference between AS and SAS games is, however, not statistically significant (two-sided WRS test, CSV1

& CSV2 $p = 0.7773$, CSV4 & CSV5 $p = 0.8899$). This implies that while the first part of Hypothesis 1 can be confirmed, the second is rejected.

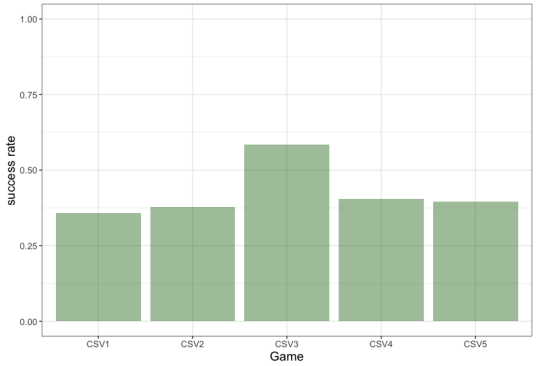
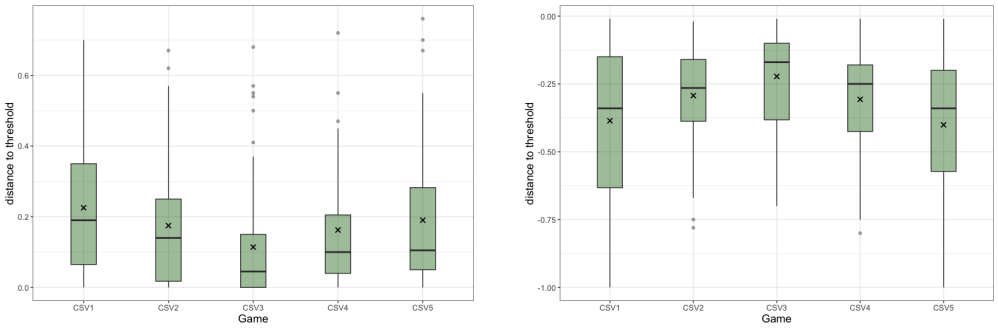


Figure 32: Success rates in CSV games

Yet, the success rate only gives an average over the whole sample, without detailing by how much the threshold was missed or overshot. Figure 33 illustrates these average distances to the threshold for the five CSV games. We can see that values for CSV2 and CSV4 (AS) are less noisy than for CSV1 and CSV5 (SAS). Interestingly, the mean overshoot for the strongly asymmetric games is 0.226 and 0.190 and the mean "miss" is by 0.386 and 0.401 respectively, all higher in absolute terms than for the asymmetric games. This indicates that even though there is no discernible difference in success rates, play in SAS games was more erratic than in AS games. This U-shape (reverse U-shape) is completed by the values for CSV3, where both overshoot and miss are closest to zero.



(a) Distance to threshold if succeeded

(b) Distance to threshold if failed

Figure 33: Success and failure in CSV games

Burden-sharing is usually defined by the the relative size of one player's contribution level to total contributions (see e.g. Waichman et al. 2021). However, in our quadratic cost setup, the actual burden a player faces is the contribution cost, the share of which does not correspond to the contribution share as in a linear setting. We therefore define the burden-share of player i as the relative share of their costs

to total costs as given by:

$$\text{share}_i = \frac{c(x_i)}{c(x_i) + c(x_j)} = \frac{5x_i^2}{5x_i^2 + 5x_j^2} \tag{14}$$

Figure 34 illustrates the distribution of cost shares across the five games. The mean contribution share in the symmetric game is significantly different to the asymmetric and strongly asymmetric games (pairwise one-sided WRS tests, highest $p = 0.0004$) whereas the means between CSV1 and CSV2 as well as CSV3 and CSV4 do not differ significantly. However, the cost shares differ with respect to their spread: in strongly asymmetric games, contribution shares have a higher variance, also indicated visually by the wider box.

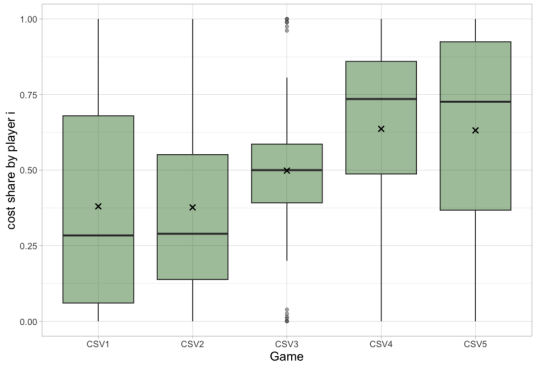


Figure 34: Distribution of cost shares in CSV games

We can also compare the cost shares between successful and unsuccessful games, plotting them against the cost shares corresponding to the fair and efficient focal points, as defined in Section 3.2.1. Figure 35 illustrates this for all five CSV games. For the symmetric game CSV3 (see Figure 35c) we can see that the burden share between the successful and unsuccessful groups was practically identical (means 0.4977 and 0.4992), both at the focal points of efficiency and fairness. The main difference between the success and failure groups, however, lies in the spread of the contribution shares, with the standard deviation being much higher for failed groups (see sd_s for successful and sd_f for failed groups). This implies that while on average the contribution share was fair and efficient, individual contribution levels were more erratic, leading to threshold misses.

A similar picture follows for asymmetric and strongly asymmetric games, that is, the mean contribution share does not differ largely between successful and failed groups, whereas the variance does. The most striking difference between asymmetric games CSV2 and CS4 and strongly asymmetric games CSV1 and CSV5 lies in the fact that in the former group, cost sharing is too equal compared to what would be both fair or efficient, whereas in the latter group, cost shares lie in between the two focal points. This implies that players seem to play both types of games very similarly, which leads to especially unequal burden-sharing in the case of strongly asymmetric games. Note that the efficient focal

point for SAS games implies that the "richer" player receives the full surplus from cooperation, and we see that on average, poor players contribute even beyond this point.

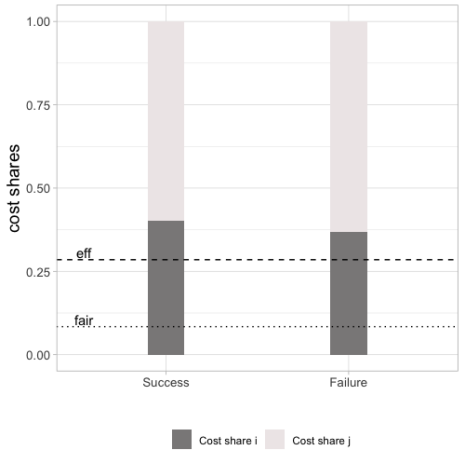
Asymmetric Games

The treatment type of asymmetric games is characterised by the existence of an equilibrium interval in which the equal split allocation is contained. Henceforth we will divide them into four subgroups, keeping the preference parameter of player i fixed and analyse contribution levels and success rates accordingly. Table 8 provides summary statistics.

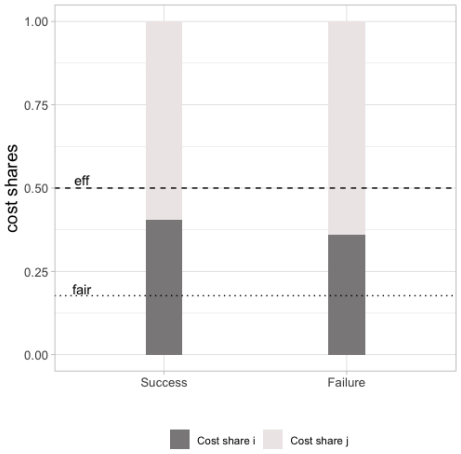
Table 8: Summary statistics asymmetric treatments

Game	Type	(θ_i, θ_j)	mean x_i	median x_i	sd x_i
AS1	AS	(1.25, 2.5)	0.3717	0.4000	0.1960
AS2	AS	(1.25, 3.75)	0.3663	0.3400	0.2277
AS3	AS	(1.25, 4.25)	0.3524	0.3100	0.2190
AS4	AS	(2.5, 1.25)	0.4507	0.5000	0.2340
AS5	AS	(2.5, 3.75)	0.4567	0.4600	0.2054
AS6	AS	(2.5, 4.25)	0.4154	0.4000	0.1756
AS7	AS	(3.75, 1.25)	0.5173	0.5100	0.2308
AS8	AS	(3.75, 2.5)	0.5025	0.5000	0.2014
AS9	AS	(3.75, 4.25)	0.4924	0.5000	0.1820
AS10	AS	(4.25, 1.25)	0.5331	0.5450	0.2323
AS11	AS	(4.25, 2.5)	0.5374	0.5450	0.1984
AS12	AS	(4.25, 3.75)	0.5020	0.5000	0.1686

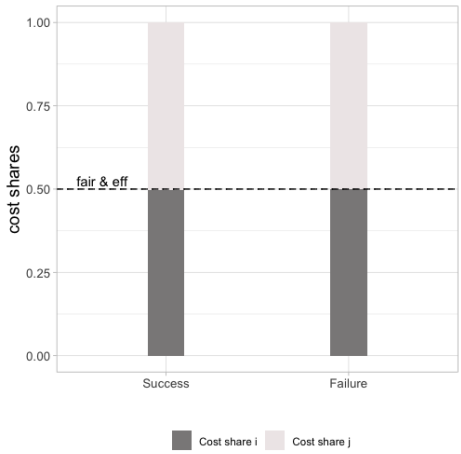
Figure 36a illustrates contribution levels, which we would expect to decrease from both AS1–AS3 and from AS4–AS6. Mean contributions do not significantly decrease from AS1–AS3 (pairwise one-sided WRS, lowest $p = 0.076$), where average contributions in AS6, which is the most asymmetric game in this subgroup, is significantly lower than in AS4 and AS5 (pairwise one-sided WRS, highest $p = 0.044$). Figure 36b analogously depicts the third and fourth subgroup, where we would expect contributions to decrease from AS7–AS9 as well as from AS10–AS12. The only statistically signifcant mean difference is between AS7 and AS9 (one-sided WRS, $p = 0.034$). Note that the most asymmetric games, that is AS3 and AS10, have among the highest contribution spreads, mirroring the more erratic play as asymmetry is increased.



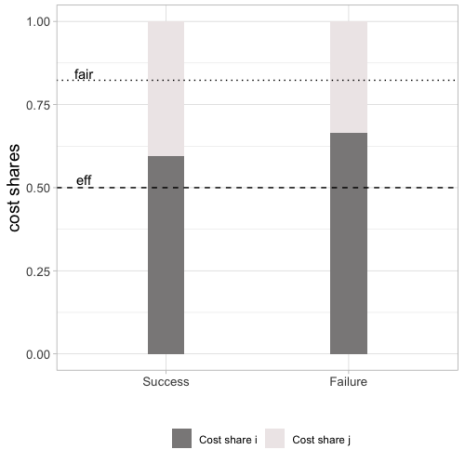
(a) CSV1: $sd_s = 0.2605$, $sd_f = 0.3907$



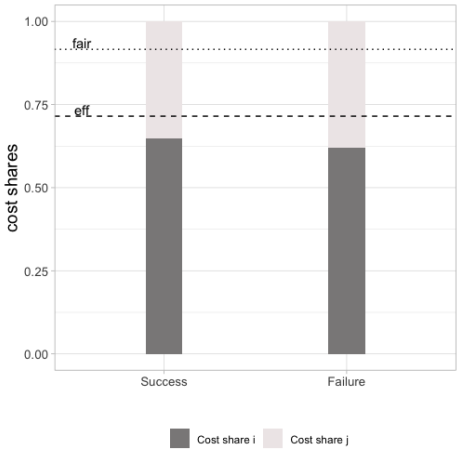
(b) CSV2: $sd_s = 0.2347$, $sd_f = 0.3488$



(c) CSV3: $sd_s = 0.1195$, $sd_f = 0.3004$

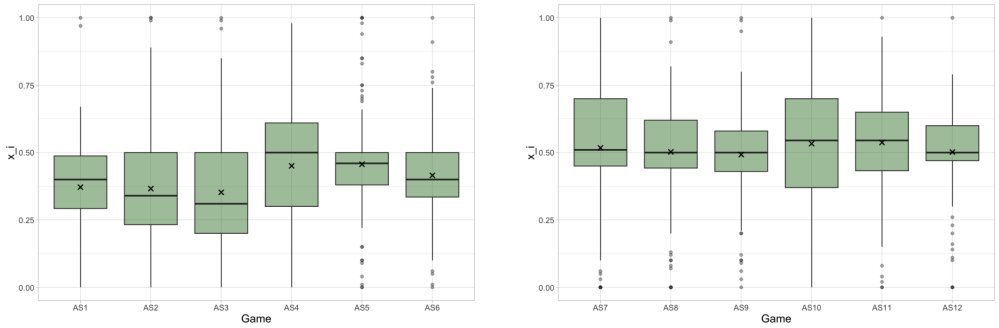


(d) CSV4: $sd_s = 0.2309$, $sd_f = 0.3483$



(e) CSV5: $sd_s = 0.2523$, $sd_f = 0.3857$

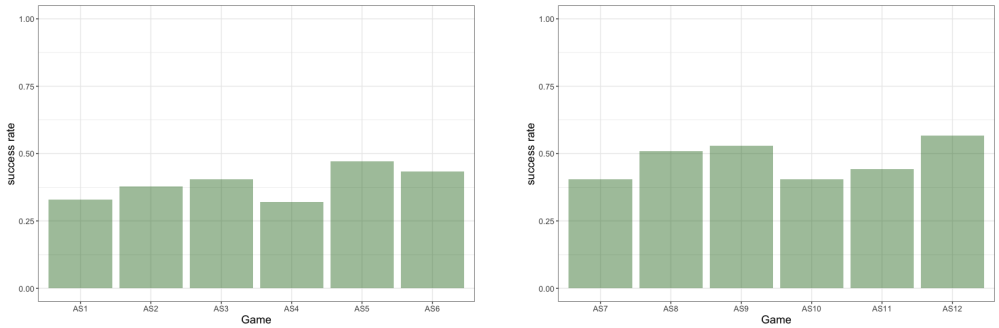
Figure 35: Burden sharing in success vs. failure games



(a) AS1–AS6 **(b)** AS7–AS12

Figure 36: Contribution distribution in asymmetric games

When looking at success rates, we have to consider two forces at play: the social value of the public good and the degree of asymmetry. We hypothesise that increasing the first should facilitate provision and increasing the latter is expected to hamper success. In the first subgroup, illustrated in Figure 37a, the two forces counteract each other (higher stakes combined with higher asymmetry) and therefore, it is unclear which effect will prevail, which shows in no statistically significant differences in success rates within games AS1–AS3. In the second subgroup, the success rate of AS4 is significantly lower compared to the other two (pairwise one-sided WRS, highest $p = 0.045$), which is unsurprising due to the low stakes. The difference between AS5 and AS6 is not significant, but the decreasing success rate hints at the existence of a harmful effect of asymmetry. The other two subgroups are depicted in Figure 37b. The success rates are significantly different between AS7 and AS9 (one-sided WRS, $p = 0.037$) and between AS12 and the other two (pairwise one-sided WRS, highest $p = 0.037$), the latter difference confirming that a simultaneous increase in stakes and decrease in asymmetry facilitates provision.



(a) AS1–AS6 **(b)** AS7–AS12

Figure 37: Success rates in asymmetric games

Strongly Asymmetric Games

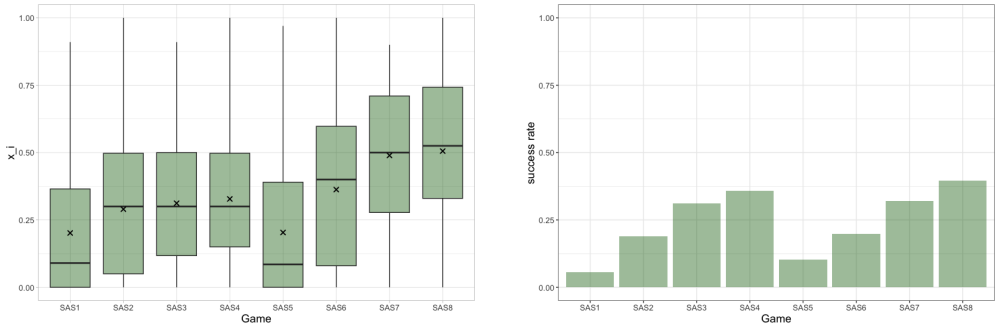
By definition, strongly asymmetric games are those in which one of the two players has a preference parameter of 0.75, meaning that the equal split allocation is not part of the equilibrium interval. Therefore, in these games, in order to reach the threshold, the player with the higher preference parameter would have to step up and contribute significantly more than half of the threshold. Note that in SAS1 and SAS5 provision is never efficient, since the "richer" player does not have sufficiently high benefits in order to contribute enough to reach the threshold in a way that both players have a non-negative payoff. Table 9 provides summary statistics for the two subgroups of this treatment type, where in SAS1–SAS4 asymmetry is against player i and in SAS5–SAS8 asymmetry is in favour of player i. Note that the efficient contribution split in strongly asymmetric games would be 0.387/0.613 for the poorer and richer player respectively.

Table 9: Summary statistics for strongly asymmetric treatments

Game	Type	(θ_i, θ_j)	mean x_i	median x_i	sd x_i
SAS1	SAS	(0.75, 1.25)	0.2010	0.0900	0.2379
SAS2	SAS	(0.75, 2.5)	0.2899	0.3000	0.2525
SAS3	SAS	(0.75, 3.75)	0.3120	0.3000	0.2295
SAS4	SAS	(0.75, 4.25)	0.3279	0.3000	0.2597
SAS5	SAS	(1.25, 0.75)	0.2034	0.0850	0.2539
SAS6	SAS	(2.5, 0.75)	0.3626	0.4000	0.2799
SAS7	SAS	(3.75, 0.75)	0.4893	0.5000	0.2628
SAS8	SAS	(4.25, 0.75)	0.5250	0.5055	0.2833

We can see that while means are significantly lower in SAS1 and SAS5 compared to the other games, average contributions are clearly positive, while median contributions are significantly lower, implying that this relatively high average is driven by few players who contributed highly above their means. The standard deviation is substantial across all 8 games. We can also see that in the game in which asymmetry is most in favour of player i, that is SAS8, they on average only contribute slightly above the equal split, which is not enough to ensure a provision combined with a non-negative payoff by their co-player. Figure 38a illustrates these facts.

Figure 38b picture success rates across the 8 games. Finally, note that the two groups SAS1–SAS4 (asymmetry against i) and SAS5–SAS8 (asymmetry in favour of i) consist of pairwise mirror games and therefore illustrate the same trend: keeping one preference parameter at 0.75, increasing stakes similarly lift the success rate.



(a) Contribution distribution **(b)** Success rates

Figure 38: Contribution and success rates in strongly asymmetric games

3.5 Conclusion

This paper discusses a one-shot threshold public goods game played in groups of two, in which players have varying preferences for the public good. This gives rise to three treatment types: symmetric, asymmetric and strongly asymmetric games, the last being characterised by the fact that the equal split allocation is not feasible. For sufficiently high preference parameters, an equilibrium interval emerges on the threshold line. Two potential focal equilibrium points are discussed: the efficient equilibrium ensures minimised contribution costs across players, whereas the fair equilibrium equally splits the surplus generated from provision. Furthermore, if on the equilibrium interval, the allocation in which both players contribute half of the threshold level, is expected to be a focal point.

Our specific experimental specification implies that in symmetric treatments, all three focal points of equal split, efficiency and fairness coincide. In asymmetric treatments, the fair focal point diverges and furthermore for strongly asymmetric treatments, the equal split is no longer an equilibrium. We thus hypothesise that provision success decreases in the degree of asymmetry. Indeed we find that groups with symmetric players have a significantly higher success rate in reaching the threshold than other treatment types. Surprisingly, the specific degree of asymmetry has no effect on average provision success, yet when considering the distribution of contributions, we see that volatility increases with stronger asymmetry. This also leads to the fact that successful groups on average overshoot the threshold by most in strongly asymmetric games, and analogously miss the threshold most clearly in failed groups.

Within a subgroup of treatments in which the social value of the public good is constant, burden-sharing between the two players is analysed. Results suggest that while experimental subjects notice and act on asymmetry, the degree of asymmetry is less salient than anticipated. This shows in relatively stable burden-sharing across a variety of asymmetric specifications, resulting in the fact that "poor" players tend to contribute too much and "rich" players too little compared to both the fair and efficient benchmark. This result hints at a potentially biased perception of asymmetries, both from the

perspective of the advantaged and disadvantaged players, leading to both relatively low success rates and unjust cost sharing in asymmetric specifications. Re-contextualising this finding to the motivating example of international cooperation on climate change mitigation, this would imply that wealthier nations might underestimate the degree to which their contributions have to exceed that of poorer countries when aiming for either efficient or fair outcomes.

Acknowledgements

I would like to thank Jean-Michel Benkert, Igor Letina, Marc Möller, Anna Schmid, Alessandro Tavoni, Ralph Winkler and seminar participants at the University of Bern for valuable comments. Special thanks to Amil Redja for the excellent research assistance. This research has received funding from the SNSF grant 189163.

Appendix

A Specifications

Table 10: Game Specifications

Game	Type	Name	(θ_i, θ_j)	Social value	Efficient x_i	Fair x_i
1	S	S1	(0.75, 0.75)	1.5	0	0
2	SAS	SAS1	(0.75, 1.25)	2	0	0
3	SAS	SAS2	(0.75, 2.5)	3.25	0.387	0.340
4	SAS	SAS3	(0.75, 3.75)	4.5	0.387	0.261
5	SAS	CSV1/SAS4	(0.75, 4.25)	5	0.387	0.233
6	SAS	SAS5	(1.25, 0.75)	2	0	0
7	S	S2	(1.25, 1.25)	2.5	0 / 0.5	0.5
8	AS	AS1	(1.25, 2.5)	3.75	0.5	0.396
9	AS	CSV2/AS2	(1.25, 3.75)	5	0.5	0.317
10	AS	AS3	(1.25, 4.25)	5.5	0.5	0.289
11	SAS	SAS6	(2.5, 0.75)	3.25	0.613	0.660
12	AS	AS4	(2.5, 1.25)	3.75	0.5	0.604
13	S	S3/CSV3	(2.5, 2.5)	5	0.5	0.5
14	AS	AS7	(2.5, 3.75)	6.25	0.5	0.421
15	AS	AS8	(2.5, 4.25)	6.75	0.5	0.393
16	SAS	SAS7	(3.75, 0.75)	4.5	0.613	0.739
17	AS	CSV4/AS5	(3.75, 1.25)	5	0.5	0.683
18	AS	AS9	(3.75, 2.5)	6.25	0.5	0.579
19	S	S4	(3.75, 3.75)	7.5	0.5	0.500
20	AS	AS11	(3.75, 4.25)	8	0.5	0.472
21	SAS	CSV5/SAS8	(4.25, 0.75)	5	0.613	0.767
22	AS	AS6	(4.25, 1.25)	5.5	0.5	0.711
23	AS	AS10	(4.25, 2.5)	6.75	0.5	0.607
24	AS	AS12	(4.25, 3.75)	8	0.5	0.528
25	S	S5	(4.25, 4.25)	8.5	0.5	0.5

B Subject Pool Demographics

Question: *Was the experiment difficult to understand?*

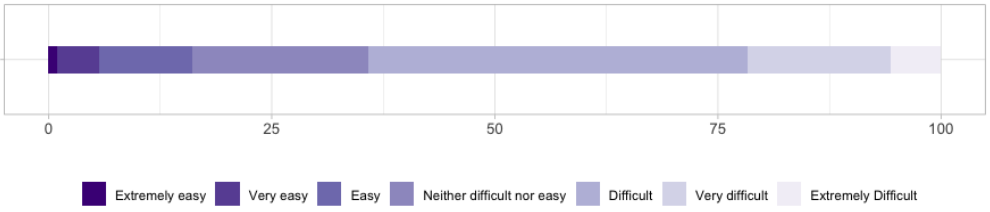


Figure 39: Difficulty

Question: *Are you generally a person who is fully prepared to take risks (risk prone) or do you try to avoid taking risks (risk averse)?*

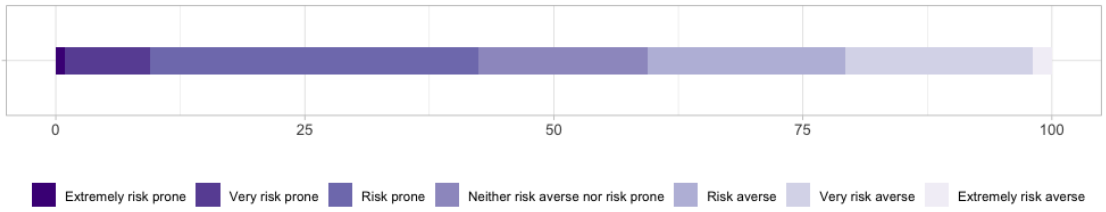


Figure 40: Risk attitude

Question: *Generally speaking, how often do you trust others?*

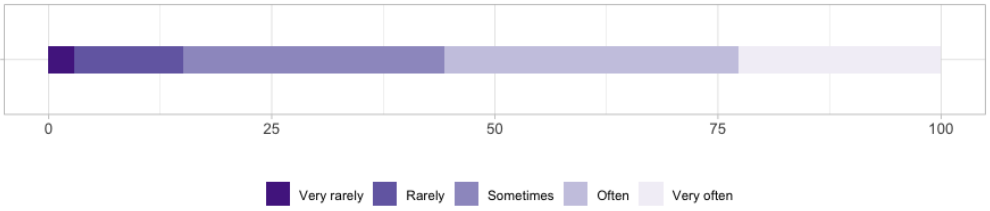


Figure 41: Trust

Choices "never" and "always" were never selected.

Question: *How good are you at working with fractions (e.g. "one fifth of something") or percentages (e.g. "20% of something")?*

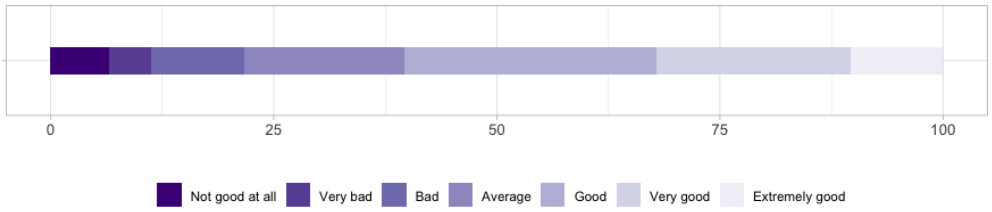


Figure 42: Fraction skills

Question: *Have you ever donated money or goods to a charitable organisation? If yes, how frequently?*

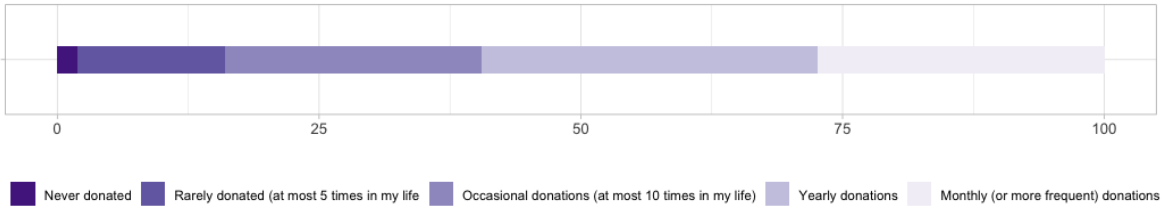


Figure 43: Donations

Question: *Which of the following guiding principles best describes your understanding of fairness in the context of the experiment you took part in (the 25 rounds you played before)?*

Answers:

(a) The player with the highest benefit from contributing to the project should invest more in it, such that payoffs are roughly the same for both players.

(b) Both players should choose 50%, irrespective of relative benefit numbers.

(c) Both players should choose 0%, irrespective of relative benefit numbers.

(d) Both players should choose what's in their own best interest, i.e. maximizes own payoffs.

(e) Players should choose the values of X_1 and X_2 that maximize the joint payoffs (i.e. the sum of the payoff from player 1 and player 2)

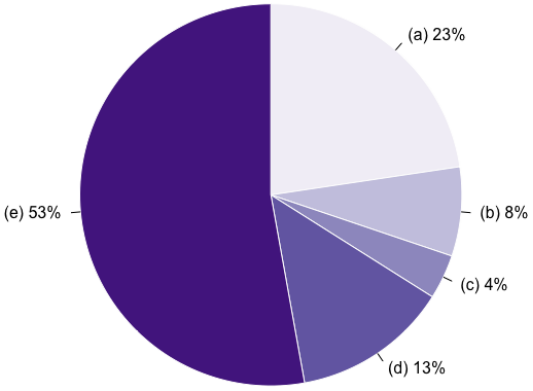


Figure 44: Understanding of fairness

Question: *Assume that benefit numbers are such that joint payoff can be positive. Which of the following statements do you agree with most?*

Answers:

(a) Players should choose contributions such that both players have a positive payoff.

(b) Players should prioritize reaching the threshold, irrespective of whether individual payoffs are positive.

(c) Both players should contribute their fair share in order to reach the threshold.

(d) The threshold should be met with precision such as to not waste any investments.

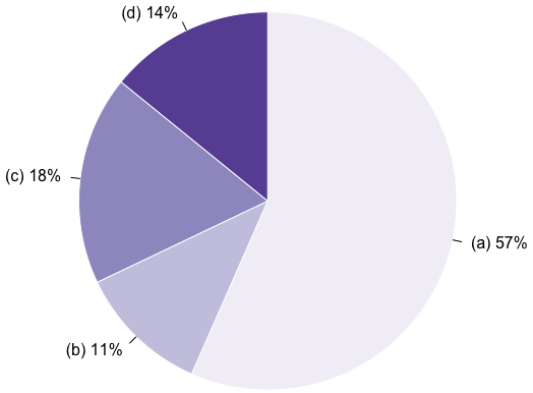


Figure 45: Statements

Question: *Please pick the MOST difficult aspect of the experiment.*

Answers:

(a) None

(b) Using the payoff calculator

(c) Guessing contribution X_2

(d) Choosing contribution X_1

(e) Grasping the differences between rounds.

(f) Something else

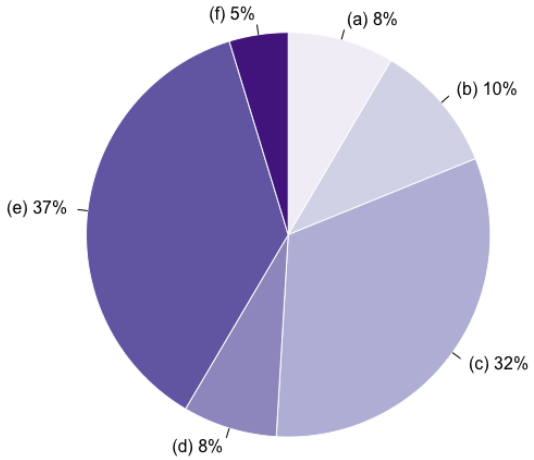


Figure 46: Most difficult aspect

Question: *What was the MOST important rationale for your decisions during the experiment?*

Answers:

(a) Monetary self-interest (i.e. maximising own income)

(b) Group efficiency (i.e. maximising joint income of both players)

(c) Minimise time spent on the task

(d) Avoiding risk (i.e. avoiding being the "sucker" who contributes a high fraction when co-player contributes little, potentially risking not reaching the threshold)

(e) Reciprocity (i.e. contributing a similar fraction to the one that the co-player was expected to contribute)

(f) Outperforming the co-player (i.e. earning a higher income than s/he)

(g) Contributing a bit extra as a precaution, such as to increase the likelihood of reaching the threshold

(h) Other

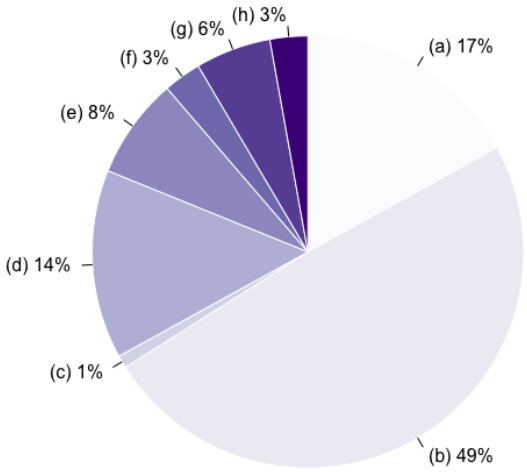


Figure 47: Most important rationale

C Further Results

Sequence effects

Note that for Figures 25a and 48, the games for which a clear fair or efficient allocation cannot be defined, have been removed from the dataset, which are games $1, 2, 6, 7$ for *fair* and game 7 for *efficient*.

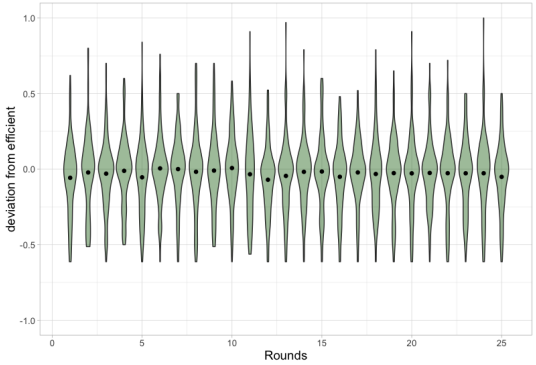


Figure 48: Deviation from efficient

References

Bagnoli, M. and B. L. Lipman (1989). Provision of public goods: Fully implementing the core through private contributions. *Review of Economic Studies 56*(4), 583–601.

Barrett, S. and A. Dannenberg (2012). Climate negotiations under scientific uncertainty. *Proceedings of the National Academy of Sciences 109*(43), 17372–17376.

Burton-Chellew, M. N., R. M. May, and S. A. West (2013). Combined inequality in wealth and risk leads to disaster in the climate change game. *Climatic change 120*, 815–830.

Croson, R. T. and M. B. Marks (2000). Step returns in threshold public goods: A meta-and experimental analysis. *Experimental Economics 2*(3), 239–259.

Dannenberg, A., A. Löschel, G. Paolacci, C. Reif, and A. Tavoni (2015). On the provision of public goods with probabilistic and ambiguous thresholds. *Environmental and Resource Economics 61*(3), 365–383.

Feige, C., K.-M. Ehrhart, and J. Krämer (2018). Climate negotiations in the lab: A threshold public goods game with heterogeneous contributions costs and non-binding voting. *Environmental and Resource Economics 70*, 343–362.

Fischbacher, U., S. Schudy, and S. Teyssier (2014). Heterogeneous reactions to heterogeneity in returns from public goods. *Social Choice and Welfare 43*, 195–217.

Gavrilets, S. (2015). Collective action problem in heterogeneous groups. *Philosophical Transactions of the Royal Society B: Biological Sciences 370*(1683), 20150016.

IPCC (2022). Climate Change 2022: Mitigation of Climate Change. Contribution of Working Group iii to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. *[P.R. Shukla, J. Skea, R. Slade, A. Al Khourdajie, R. van Diemen, D. McCollum, M. Pathak, S. Some, P. Vyas, R. Fradera, M. Belkacemi, A. Hasija, G. Lisboa, S. Luz, J. Malley, (eds.)]*.

Klinsky, S., T. Roberts, S. Huq, C. Okereke, P. Newell, P. Dauvergne, K. O'Brien, H. Schroeder, P. Tschakert, J. Clapp, et al. (2017). Why equity is fundamental in climate change policy research. *Global Environmental Change 44*, 170–173.

Lange, A., A. Löschel, C. Vogt, and A. Ziegler (2010). On the self-interested use of equity in international climate negotiations. *European Economic Review 54*(3), 359–375.

McBride, M. (2010). Threshold uncertainty in discrete public good games: an experimental study. *Economics of Governance 11*(1), 77–99.

McGinty, M. and G. Milam (2013). Public goods provision by asymmetric agents: experimental evidence. *Social Choice and Welfare 40*, 1159–1177.

Milinski, M., R. D. Sommerfeld, H.-J. Krambeck, F. A. Reed, and J. Marotzke (2008). The collective-risk social dilemma and the prevention of simulated dangerous climate change. *Proceedings of the National Academy of Sciences 105*(7), 2291–2294.

Palfrey, T. R. and H. Rosenthal (1984). Participation and the provision of discrete public goods: a strategic analysis. *Journal of Public Economics 24*(2), 171–193.

Rapoport, A. and R. Suleiman (1993). Incremental contribution in step-level public goods games with asymmetric players. *Organizational behavior and human decision processes 55*(2), 171–194.

Tavoni, A., A. Dannenberg, G. Kallis, and A. Löschel (2011). Inequality, communication, and the avoidance of disastrous climate change in a public goods game. *Proceedings of the National Academy of Sciences 108*(29), 11825–11829.

UNEP (2022). Emissions gap report 2022: The closing window – climate crisis calls for rapid transformation of societies. *https://www.unep.org/emissions-gap-report-2022*.

Waichman, I., T. Requate, M. Karde, and M. Milinski (2021). Challenging conventional wisdom: Experimental evidence on heterogeneity and coordination in avoiding a collective catastrophic event. *Journal of Environmental Economics and Management 109*, 102502.

Selbstständigkeitserklärung

Ich erkläre hiermit, dass ich diese Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen benutzt habe. Alle Stellen, die wörtlich oder sinngemäss aus Quellen entnommen wurden, habe ich als solche gekennzeichnet. Mir ist bekannt, dass andernfalls der Senat gemäss Artikel 36 Absatz 1 Buchstabe o des Gesetzes vom 5. September 1996 über die Universität zum Entzug des aufgrund dieser Arbeit verliehenen Titels berechtigt ist.

Bern / 03. Februar 2023

Sarah Spycher
Name