Present Bias in Politics and Self–Committing Treaties *

Bård Harstad Anke Kessler

March 1, 2025

Abstract

We study how international environmental agreements can take advantage of domestic time-inconsistency problems. Policymakers often prefer future policies to be sustainable, but are tempted to invest less when being in office. We find the equilibrium number of signatory countries to be higher than when preferences are time consistent, especially when the political environment is unstable and polarized and the international spillovers are limited. This model also explains participation in treaties whose mandates do not vary with the coalition size and why the coalition will not unravel if, for example, the US exits the Paris Agreement.

Keywords: international treaties, time inconsistency, self-commitment, environmental policy

JEL classification: Q54, Q58, H87, F53,

*Stanford University and Simon Fraser University, respectively, with emails harstad@stanford.edu and akessler@sfu.ca. Max Wosnitza has been an excellent research assistant. Harstad's part of the research received funding from the European Research Council, ERC GA no. 101055353. Kessler acknowledges financial support from the Chair's Research Grant, Simon Fraser University. All errors are our own.

1 Introduction

Policy preferences of governments typically fail to be dynamically consistent. In fact, optimal policies are necessarily time inconsistent if policymakers rotate being in office.[1] In such a situation, contemporary governments always hope that future governments will act sustainably, for example, preferring to defer costly action to the future themselves.

Environmental policies are textbook examples of decision problems that suffer from time inconsistency or "present bias." Pro-environment action generally involves future gain at immediate expense: most policies take the form of investments, trading off costs today with benefits tomorrow or as distant as several generations into the future. Examples are emission reduction, conservation of natural habitat, protection of species, and conservation of natural resources.

In this paper, we explore the consequences of time-inconsistent preferences for countries' incentives to sign international environmental agreements (IEAs) and to uphold the treaty over time when compliance is endogenous. We build a simple model of treaty formation and enforcement that captures the essential elements of an environmental policy decision and policymakers who rotate being in office. We show that while time inconsistency leads to a political failure for domestic politics and inefficiencies, the desire to attempt to tie the hands of future policymakers is a weakness that international treaties can take advantage of. That is, international cooperation can be facilitated by domestic political failure.

Three results emerge from our analysis. First, we demonstrate that the weakness in domestic politics increases the scope for participation: the equilibrium number of signatories is larger than when preferences are time consistent and increases as the domestic bias toward the present worsens. Second, domestic policy failure can strengthen compliance if enforcement is not guaranteed, i.e., if the treaty needs to be sustained by the credible threat of punishment among its member countries. Intuitively, countries can be more motivated to comply with an agreement when preferences are time inconsistent because they suffer more from the business-as-usual equilibrium they expect when being outside of the agreement. But retaliations come in the future, so too much present-bias leads to defections.

Third, since one motivation for countries to participate is to lock in domestic policies rather than to strengthen the contribution of others, the theory can provide a rationale for treaties or "conventions" that specify mandates irrespective of the coalition size. Specifically, we show that "conventions" with fixed mandates that are initially negotiated but remain in force even as more countries sign on to the treaty are attractive and can be sustained in equilibrium, unlike in the standard model.

In sum, the political turnover can be necessary to create the weakness that an IEA can exploit, and that can make treaties large and robust. Elections with backward-looking voters weaken but do not change the results, qualitatively.

These results bear empirical relevance in several respects. First, the theory illustrates that in a world

[1] It is well known that political turnover leads to time inconsistency (e.g., Persson and Svensson, 1989; Alesina and Tabellini, 1990; Tabellini, 1991). The fact that policymaker rotation leads to time-inconsistent preferences follows from Amador (2003), Chatterjee and Eyigungor (2016), and Harstad (2020).

of domestic present bias, international agreements have the added benefit of tying the hands of future home governments with regard to *domestic* policy issues. This benefit increases countries' incentives to join international agreements and is larger the more pronounced the domestic policy problem is. In our context of environmental policies, this means that the participation and compliance scope of IEAs widens further in cases where the environmental issue to be addressed with the policy has a significant local component. The theory thus provides insights to why we observe the number and scope of environmental agreements in effect today. The International Environmental Agreements Database Project lists over 3,800 such agreements, of which roughly 40 percent are multilateral. In 2025 alone, the Geneva Environmental Network names 18 large environmental conferences to facilitate international negotiations and compliance on environmental policy, many of which are held within existing frameworks of international cooperation and arguably have national (local) policy components.[2]

Second, the model can provide an explanation for international agreements that are negotiated and ratified by an initial group of signatories, but are expanded by more countries signing onto the agreement over time *even though the original text of the treaty remains (largely) unaltered.* Agreements that are not (re)negotiated to internalize spillovers on additional members as they join cannot be explained by traditional theories, where the only benefit from signing is that other members' commitments increase to account for the new treaty member, and thus would have to be renegotiated every time a new country joins. In our model, in contrast, countries may want to join even if new members do not affect policy, simply because current governments perceive an additional gain from committing future domestic policymakers to the given (treaty) policy. The majority of IEAs have, in fact, been of this nature.[3] Prominent examples are the Ramsar Convention on Wetlands, the Aaarhus Convention on Public Access to Information and Justice regarding the Environment, the UN Biodiversity Convention, the Convention on the Conservation of Migratory Species of Wild Animals, and the Convention for the Protection and Conservation of Sea Turtles, among others (see Table 1 in the Appendix for details and more examples). In line with our rationale, countries' commitments in many of those treaties arguably affect their own residents directly; in fact many conventions explicitly include domestic policy objectives such as local habitat protection, national accountability of environmental policy towards citizens, or reduction of pesticide and PCB use.

Lastly, the added benefit from (self-)commitment on domestic policy issues also accounts for more stability in agreements than what the standard theory would predict. In fact, in conventions, the contributions of other signatories are immune to defection or exit of another member state. By construction, if a country reneges on its pledges or withdraws from the treaty, then the other countries will not alter their contributions; they are still motivated to participate and comply in order to deal with their own domestic time-inconsistency problem. The increased stability implies, for example, that the US withdrawal from the Paris Agreement would not necessarily lead to reduced participation or contributions from the other countries. In the traditional theory, in contrast, without those types of reactions, there would be no hope of inducing countries to participate and comply in the first place. In this sense, our analysis permits a more optimistic point of view regarding

[2]See https://www.genevaenvironmentnetwork.org/resources/updates/news-dates-for-major-environmental-negotiations/. Those include conferences of parties (COPs) for the UN Climate Change Conference, the UN Ocean Conference, the UN Biodiversity Conference, as well as annual meetings for over a dozen international conventions (see below and Table A1 in Appendix A for examples).

[3]To the extent that these treaties are expanded in scope, the additional commitments of member countries are generally unrelated to new additional parties joining.

the future of the Paris Agreement.

Literature. By combining time inconsistency and the formation of environmental agreements, we contribute to two strands of literature. First, we draw on a long tradition of political economy models studying time inconsistency and strategic commitments – going back to Kydland and Prescott (1977). Fischer (1980) explained that even when capital taxes should be low to motivate savings, policymakers would be tempted to raise the taxes after the investments are sunk. In this setting, international cooperation can be harmful because it eliminates the competition for capital among nations that could motivate low taxes despite the time-inconsistency problem; see Rogoff (1985), van der Ploeg (1988), and Kehoe (1989).

The idea that international treaties can alleviate domestic time-inconsistency problems has primarily been explored in the area of international trade. Staiger and Tabellini (1987), Matsuyama (1990), and Maggi and Rodruiguez (1998, 2007) all highlight how time-inconsistency challenges in domestic trade policy that arise for economic or political reasons are mitigated by binding agreements that are facilitated through international institutions such as the GATT/WTO. Staiger and Tabellini (1999) present evidence on the effectiveness of this strategy by showing that GATT rules helped the US government to make domestic trade policy commitments that it could not have made otherwise. Conconi and Perroni (2009) study the relationship between domestic and international policy credibility in a general repeated game framework, where deviations are followed by noncooperation of other domestic and international players. They also consider an application to environmental policy, where the reason why the domestic choice of emission tax fails to be time-consistent is that once firms have already invested in green technology, lowering emission taxes will reduce the distributional burden without altering incentives.

Second, we contribute to the body of research that focuses on the equilibrium size of IEAs. The typical finding in this literature is that fully enforceable international agreements are incentive-compatible only if they involve a very small number of countries (Hoel, 1992; Carraro and Siniscalco, 1993; Barrett, 1994).[4] This prediction clashes with the observation that real-world coalitions are often quite large, leading to the "paradox of international environmental agreements" (Kolstad and Toman, 2005; see also Nordhaus, 2015). Among the explanations that have been proposed to explain this puzzle, our theory complements that of Battaglini and Harstad (2020), where incumbents sign weak treaties in order to influence the probability of winning the next election.[5] Marchiori et al. (2017) study how domestic lobby groups affect the size of stable IEAs, and show that the government's desire to improve its bargaining position with respect to strong anti-emission lobbies may increase its incentives to sign an IEA.

Our model differs from those above and our analysis contributes to both literatures by deriving the equilibrium coalition size as a function of the time-inconsistency problem. We study participation as well as compliance, agreements where countries' commitment levels depend on coalition size as well as agreements where they do not, and we highlight when and how the treaty can exploit the domestic time-inconsistency problem.

Outline. Section 2 presents a simple model, from which we share our main results in Section 3. Section

[4]See also Barrett (2005) and Aldy and Stavins (2009) for a survey and further references.

[5]In an extension of Battaglini and Harstad, Spycher (2024) a model where a "brown" government negotiates an unpopular IEA that a possible green successor would ratify but they would not, thereby strategically increasing its reelection chances.

4 integrates elections and backward-looking voters, followed by concluding remarks in Section 5. Auxiliary results and tables are in Appendix A. All proofs are in Appendix B.

2 The Model

2.1. The Stage Game

Consider a set N of n countries contributing to a local public good or, equivalently, a local public bad. For consistency, we will refer to emissions as a proxy to a local public bad and abatements as a proxy to a local public good. Time is discrete and the horizon is infinite. Current emissions increase the pollution stock in the future. Let $G_{i,t}$ denote the stock of pollution in country i at time t and let $1 - q_G \in [0, 1]$ measure the fraction of G that "depreciates" every period.

The stock of pollution in i can also depend on the emissions of other countries $j \neq i$. Today's collective emissions together with the current stock of pollution determine tomorrow's pollution stock as follows:

$$G_{i,t+1} = q_G G_{i,t} + \gamma g_{i,t} + \epsilon \sum_{j \in N \setminus i} g_{j,t}. \tag{1}$$

Every unit of i's emission leads to $\gamma > 0$ units of local pollution and $\epsilon \geq 0$ units of pollution in every other country. For climate change, $\gamma = \epsilon$, but for regional problems, $\gamma > \epsilon$.

The benefit of emissions accrues through consumption of a dirty good (e.g., energy). The harm of emissions is that they accumulate in $G_{i,t}$. To help derive closed-form solutions, we assume a quadratic functional form for the per-period payoff from emissions and constant marginal harm:

$$u_{i,t} = -\frac{b}{2}\left(g^*_{i,t} - g_{i,t}\right)^2 - cG_{i,t}. \tag{2}$$

Here, $g^*_{i,t}$ is a parameter representing a country's bliss point, which measures the ideal amount of emissions if there were no concern for pollution: due to implicit cost associated with generating, transporting, or consuming the dirty good that causes emissions, no country would emit above its bliss point. The parameter c is the cost of the public bad, and b measures the benefit of being close the bliss point. Because we do not require $g_{i,t}$ to be positive, the model allows for an alternative interpretation where $-G_{i,t}$ is a local public good such as a natural resource, the preservation of a critical ecosystem, or the protection of a species that depends on country i's conservation efforts g_i as well as on the efforts of other countries that spill over into i, e.g., if the resource or the ecosystem crosses the border or the species is migratory. Country i's contribution or investment to the local public good is (the abatement level) $a_{i,t} \equiv g^*_{i,t} - g_{i,t}$.

2.2 Dynamics and Time Inconsistency

At each time t, country i is run by party or policymaker in power, $P_{i,t}$. The costs from climate policies and climate change are drawn from the government's budget. The remaining budget is allocated according to $P_{i,t}$'s preference. We suppose that each remaining dollar has the additional value Δ for the party in

power, relative to the party not in power, which causes preferences between the current government and the opposition to diverge. Let p be the probability with which $P_{i,t}$ is in power in the future.

With these assumptions, $P_{i,t}$ would like to maximize

$$(1+\Delta)\,u_{i,t} + \sum_{\tau=t+1} (1+p\Delta)\,\delta^{\tau-t} u_{i,\tau} = \left[u_{i,t} + \sum_{\tau=t+1} \left(\frac{1+p\Delta}{1+\Delta}\right)\delta^{\tau-t} u_{i,\tau}\right](1+\Delta)\,.$$

Because the last factor, $(1+\Delta)$, is a constant, we get:

Lemma 1 *$P_{i,t}$ acts as if she maximizes*

$$u_{i,t} + \beta \sum_{\tau=t+1} \delta^{\tau-t} u_{i,\tau}, \quad \textit{where} \tag{3}$$

$$\beta \equiv \frac{1+p\Delta}{1+\Delta}. \tag{4}$$

Although we assume that the cost of climate change, $cG_{i,t}$, reduces $P_{i,t}$'s budget, the result would be similar if, instead, this cost reduced everyone's utility directly.[6] In either scenario, $P_{i,t}$'s objective is to maximize a continuation value characterized by quasi-hyperbolic discounting. When $\beta \in (0,1)$, these preferences are time inconsistent: $P_{i,t}$ wishes that $P_{i,t+1}$ would emit less, or abate more, but this plan will not be followed. For any $\beta < 1$, the next government will abate too little from the perspective of the current policymaker. The smaller is β, the larger is the disagreement between the plan that seems optimal today vs. the plan that will actually be followed.

We see that two factors determine the size of disagreement between current and future policymakers: β is small if the rotation of political power is frequent, i.e., for small values of p, and if the preferences are more polarized, in that the additional value of spending dollars when one is in power (Δ) is large. We may also interpret Δ as a measure of corruption. This way of rationalizing quasi-hyperbolic discounting in politics is in line with Amador (2003), Chatterjee and Eyigungor (2016), and Harstad (2020, 2023b).[7]

2.3. The First Best

The first best (FB) is defined as the allocation that maximizes $\sum_i \sum_{t=0} \delta^t u_{i,t}$. As is easily seen, for each country i and time $t > 0$ we must have:

$$a_{i,t} = a^{FB} \equiv \delta(\gamma + (n-1)\epsilon)C/b, \tag{5}$$

[6] In this case, $P_{i,t}$ would maximize $-\frac{b}{2}\left(g^*_{i,t} - g_{i,t}\right)^2(1+\Delta) - cG_{i,t} + \sum_{\tau=t+1}\delta^{\tau-t}\left[-\frac{b}{2}\left(g^*_{i,t} - g_{i,t}\right)^2(1+p\Delta) - cG_{i,t}\right]$, which $P_{i,t}$ can solve by maximizing $\frac{b}{2}\left(g^*_{i,t} - g_{i,t}\right)^2 - c_P G_{i,t} + \sum_{\tau=t+1}\beta\delta^{\tau-t}\left[-\frac{b}{2}\left(g^*_{i,t} - g_{i,t}\right)^2 - c_P G_{i,t}\right]$, where $c_P \equiv c/\beta$, because $P_{i,t}$ takes $G_{i,t}$ as given. So, $P_{i,t}$ would act as if she had quasi-hyperbolic discount factors in this case as well, and β would be the same as in Lemma 1. The difference to our formalization would be that β also influences $P_{i,t}$'s emphasis on future climate change costs, c_P.

[7] Harstad (2020) explains that Lemma 1 may hold also because of: intergenerational altruism (Phelps and Pollak, 1968); individual discount factors may be heterogeneous (Gollier and Zeckhouser, 2005) or uncertain (Gollier and Weitzman, 2010); or all individuals may be endowed with quasi-hyperbolic discount factors (Laibson, 1997).

where $C \equiv c/(1-\delta q_G)$ is the present-discounted harm of a unit of pollution that will slowly depreciate over time. Note that the FB requires that the $a_{i,t}$'s be identical across the countries and over time even though the bliss points $g^*_{i,t}$ vary.

2.4. Business as Usual

In every period, the policymakers simultaneously and noncooperatively set the $a_{i,t}$'s, taking the other countries' abatements as given. In the environmental literature, this scenario is often referred to as "business as usual" (BAU).[8] Because there is a large number of subgame-perfect equilibria (SPEs) in dynamic games, it is common to restrict attention to Markov-perfect equilibria (MPE) where players' strategies depend only on current stocks; they are not history-dependent. It is straightforward to show that there is a unique MPE of this game:[9]

$$a_{i,t} = a^{bau} \equiv g^*_{i,t} - g^{bau}_{i,t} \equiv \beta\delta\gamma C/b. \tag{6}$$

As in the FB, each country reduces consumption relative to the bliss level by the same amount. Compared to the FB, countries abate too little both because they do not take into account the externality $\epsilon > 0$, and because $\beta < 1$. Thus, even if $\epsilon = 0$, $P_{i,t+1}$ will abate too little (from the viewpoint of $P_{i,t}$) because $P_{i,t+1}$ will emphasize the personal expense of abatement (or the benefit from emitting) more than what earlier decision-makers found to be optimal.

In contrast, if $P_{i,t}$ could commit to the future abatement level, it would prefer:

$$a_{i,\tau} = g^*_{i,\tau} - g_{i,\tau} = \delta\gamma C/b > a^{bau}_{i,\tau}, \tag{7}$$

for every $\tau > t$, but $P_{i,\tau}$, in power at time τ, will prefer only $a^{bau}_{i,\tau}$.

3 Results on Participation and Compliance

3.1. Deep and Binding Agreements

We start by considering the standard two-stage participation game (see, for example, Barrett, 2005). At the beginning, at $t = 0$, each of the n countries simultaneously decides whether to participate in a coalition M that negotiates an agreement. While the length of the treaty could be part of the negotiations, we simplify and shorten the exposition by restricting attention to treaties that are signed for the remaining duration of the game, i.e., that are infinite.[10] Next, signatory countries $i \in M$ negotiate abatement levels $a_{i,t}$, for every $t > 0$, while every $P_{i,t}$, $i \notin M$, contributes noncooperatively and without commitment.

[8] The "business as usual" scenario is thus defined as the equilibrium in the absence of any international cooperation that induces countries to deviate from their individual bliss points with respect to emissions. Note, thought, that those bliss points not necessarily correspond to zero abatement.

[9] The MPE here is unique because the constant marginal harm from emissions implies that country i's best response to the strategies of other countries (assuming those do not depend on stocks) is independent of G_t. Thus, Markov-perfect strategies do not condition on the stock of pollution. The outcomes in the unique MPE of the infinite horizon game are also identical to the limit of the unique SPE in any finite horizon game for $T \to \infty$.

[10] Our results are unchanged if we allow for arbitrary agreement duration. For a treatment of endogenous duration T, see Battaglini and Harstad (2016), where countries can invest in technology as well as abate, making the game with an endogenous T more interesting.

The agreement is binding and compliance is supposed not to be an issue. An example of this kind of agreement would be the Kyoto Protocol where emission reduction targets were negotiated but are legally binding. Note that the symmetry in payoffs implies all countries in a treaty collectively agree on what the optimal abatement levels are. They also have the same benefits and costs of abating relative to the default (BAU) outcome.[11] Hence, every bargaining outcome that is efficient and symmetric (as long as the underlying game is symmetric) leads to:

$$a_{i,t} = a(m) = \delta(\gamma + (m-1)\epsilon)C/b, i \in M, t > 0. \tag{8}$$

A treaty that specifies abatement levels according to (8) is referred to as a *deep agreement* because the contributions internalize all spillovers on coalition members. Any treaty specifying abatement less than what is optimal for the coalition, in contrast, would be a *shallow agreement*. We discuss shallow agreements below.

The optimal abatement levels for countries outside the coalition are straightforward. Recalling that the unique BAU contributions are independent of stocks, every $P_{i,t}$, $i \notin M$, emits according to (6).

We can now analyze the initial participation stage of the game at the beginning of the game. For M^* to be an equilibrium coalition, no nonsignatory country should wish to join (external stability) and no signatory country should wish to leave (internal stability). The cost of participation in the treaty is that members must abate more than the level that would maximize their individual objectives; a benefit is that other participants will internalize the harm on one additional member. Yet, all countries outside the coalition benefit from this internalization as well – the agreement itself is a public good. The latter effect implies strong incentives to free-ride.

Proposition 1 *If $\beta = 1$, m is an equilibrium coalition size if and only if $m \leq 3$.*

(i) When $\beta < 1$, m can be larger, and m is an equilibrium coalition size if and only if:

$$m \leq 2 + \sqrt{1 + \frac{1-\beta^2}{(\epsilon/\gamma)^2}}.$$

(ii) The FB, with $m = n$, can be supported in equilibrium if and only if:

$$\beta \leq \sqrt{1 - (\epsilon/\gamma)^2[(n-2)^2 - 1]}.$$

The benchmark result that $m \leq 3$ when discounting is exponential is a long-standing result in environmental economics, and it gives rise to the "paradox of IEAs" mentioned in the Introduction: the number of signatory countries to IEAs predicted by theory is smaller than what we observe in practice. That the maximum coalition consists of three countries is a special case that arises in a model with linear-quadratic benefit functions (e.g. Hoel, 1992 or Barett, 2005). In other cases, larger coalitions are feasible (Karp and Simon,

[11] We can easily extend the model to allow for lower weights on the payoffs of others, as in Finus and Maus (2008) and Harstad (2023a), without changing the results qualitatively.

2013), but the general point remains that large coalitions tend to be unstable. Intuitively, the individual net cost of joining a treaty relative to BAU rises in the number of treaty countries because of the required internalization of spillovers to others in a deep agreement that prescribes the optimal contributions for the coalition. This effect puts strict limits on the size of voluntary coalitions.[12]

Proposition 1 shows that when policymakers' plans exhibit present bias, the coalition can be larger. The intuition is especially sharp when ϵ/γ is close to zero: Then, there are negligible externalities and the entire purpose of the IEA is to ensure that future policymakers contribute levels preferred by the current policymakers. In this case, therefore, everyone will participate.

If ϵ/γ is larger, there is a trade-off: Participation allows the incumbent to tie the hands of future policymakers, but one will also be expected to contribute more so as to take the international externalities into account. The first effect is especially large when β is small, and the time inconsistency problem is severe. In this case, the equilibrium m is large. From the perspective of other signatories, the agreement exploits a weakness (failure) in domestic politics.

It is straightforward to extend the above finding to shallow agreements that specify lower than first-best abatement levels for the coalition. The literature has long argued that shallow agreements with more limited commitments can increase voluntary participation and thus allow for larger equilibrium coalition sizes, i.e., there is a narrow-but-deep versus broad-but-shallow trade-off. It is therefore possible that total emissions in a shallow agreement are lower (Finus and Maus, 2008). Shallow treaties would, for example, endogenously arise if signatory countries weigh other countries' utilities less than their own at bargaining stage (Harstad, 2023a). As is easily seen, present bias will broaden the scope for coalition formation for shallow agreements, for the same reason as before. The equilibrium number of signatory countries would thus increase further.[13]

3.2 Self-Enforcing Agreements and the Best SPE

An essential feature of international agreements is that enforcement through sufficiently severe sanctions for noncompliance is difficult and often impossible. The question then arises of whether the countries will comply and contribute as expected. Given the incentive to free ride, it is reasonable to be concerned with the temptation to contribute less at the time when other participants are expected to deliver on their promises. After all, because decisions are made simultaneously, a country that defects will be able to enjoy the benefit from the other contributions in that period.

We start by abstracting from the participation game and instead search for conditions under which contributions are *self-enforcing*, that is, countries are incentivized to comply with their pledges because other countries might cease to cooperate otherwise. Allowing for history-dependent strategies requires us to relax the MPE equilibrium refinement. Instead of characterizing all subgame-perfect equilibria in this dynamic game, however, we will present the most efficient symmetric SPE. Specifically, we assume that the countries employ trigger strategies in that they all revert to BAU (i.e., the noncooperative MPE) with probability

[12] It should be noted, though, that counting signatories of existing treaties may oversimplify the issue. In the case of the Kyoto Protocol, for example, while some countries – notably the United States – never ratified the agreement, others like Canada ratified but later withdrew. In addition, East European signatory countries enjoyed very lax commitment and non-Annex I (developing) countries ratified without a legally binding emissions limitation target.

[13] A formal proof is available from the authors upon request.

$q \in [0,1]$ if any one of the other abated less than a in the previous period. (Thus, with probability $1-q$, the commitments are sticky and unresponsive to a country's defection.) Note that if, with probability q, countries can observe that a defection has occurred but not which country that defected, these strategies are "minmax" and they can implement the largest possible a in any SPE.[14] Thus, we may refer to the best SPE as the second best.

Proposition 2 *(i) The best SPE is:*

$$a^{SB} = \min\left\{a^{FB}, \frac{\beta\delta C}{b}\frac{\gamma\left(1+\delta\left[1-\beta\right]\right)+2q\left(n-1\right)\delta\epsilon}{1-\delta\left(1-\beta\right)}\right\}.$$

(ii) The first best is self-enforcing if $a^{SB} = a^{FB}$, which implies:

$$\frac{\epsilon}{\gamma}\left[1-\delta\left(1-\beta\left[1-2q\right]\right)\right] \leq \left[\delta\left(1+\beta\right)-1\right]\frac{1-\beta}{n-1}.$$

(iii) Otherwise, a^{SB} decreases in β iff:

$$\frac{\epsilon}{\gamma} \leq \frac{1}{\delta q\left(n-1\right)}\left(\frac{\left[1-\delta\left(1-\beta\right)\right]^2/2}{1-\delta}-1\right).$$

This inequality is more likely to hold when ϵ/γ is small and β is large. If $\beta \to 1$, it boils down to the following condition, which always holds for ϵ/γ sufficiently small as long as $\delta > 1/2$.

$$\frac{\epsilon}{\gamma} \leq \frac{\delta - 1/2}{\delta q\left(1-\delta\right)\left(n-1\right)}.$$

In this case, a decline in β, from 1, increases a^{SB} and thus the strength and the benefit of the best self-enforcing treaty. The intuition is that when β declines, the benefit from self-commitment increases (especially when γ is large relative to ϵ), and thus it becomes more important to ensure that cooperation will continue. This importance reduces the temptation to defect.

If β is very small, however, the future payoff is discounted by a lot, relative to current payoffs. In this case, a smaller β will raise the temptation to defect, and this temptation reduces a^{SB}. Thus, a^{SB} is hump-shaped in β.

3.3 Deep and Self-Enforcing Agreements

Here, we combine the insights from the previous two subsections. For a coalition of size m, (8) defines the coalition's optimal $a\left(m\right)$. We will say a treaty with coalition size m and pledges $a(m)$ is self-enforcing if the following constitutes an SPE: Every country $i \in M$ sets $a(m)$ in every period $t \geq 1$ unless one country $i \in M$ sets $a_{i,t} \neq a(m)$ in some period $t \geq 1$, in which case, with probability $q \in [0,1]$, everyone in M reverts to a^{bau}

[14] These types of self-enforcing agreements might not be renegotiation proof. Agreements between multiple parties can be supported and be renegotiation proof if the parties use trigger strategies that either punish the defector more than they punish the other parties or give the defector no bargaining power in the renegotiation game (Mailath and Samuelson, 2006).

in $t+1$ and forever after. Every country $i \notin M$ sets a^{bau} at time $t=0$ and all future periods, independent of the history of the game.

These contribution levels $a(m)$ are self-enforcing if, analoguously to Proposition 2(i),

$$a(m) \leq \frac{\beta\delta C}{b} \frac{\gamma\left(1+\delta\left(1-\beta\right)\right)+2q\left(m-1\right)\delta\epsilon}{1-\delta\left(1-\beta\right)}.$$

By combining this compliance constraint with (8), we get,

Proposition 3 *(i) Deep agreements are self-enforcing if and only if:*

$$\frac{\epsilon}{\gamma}\left(1-\delta\left[1-\beta\left(1-2q\right)\right]\right) \leq \frac{\left(\delta\left[1+\beta\right]-1\right)\left(1-\beta\right)}{m-1}.$$

(ii) When the inequality binds, it is relaxed (so that it holds for a larger set of ϵ's and γ's) by a decrease in β from 1 if and only if $\delta > 1/2$.

As for Proposition 2, the commitments are more likely to be self-enforcing if ϵ is small relative to γ. As above, a smaller $\beta < 1$ can reduce the temptation to defect because it increases the commitment value of the treaty. Once again, the capacity of a self-enforcing deep agreement to be sustainable (e.g. the threshold for ϵ/γ) is not monotone but rather is hump-shaped in β.

The combination of Proposition 1 (participation) and Proposition 3 (compliance) is illustrated in Figure 1, which is drawn for specific parameter values $q = 1/2$, $\delta = 9/10$, and $\epsilon/\gamma = 1/10$. The maximum coalition size m as a function of β (from Proposition 1) is the decreasing curve in the figure, while the compliance constraint (from Proposition 3) is given by the hump-shaped curve. If β decreases from 1, both constraints are relaxed and the coalition can be larger.

3.4. Conventions

In the traditional literature for environmental coalitions (and in the previous Sections 3.1 – 3.3), the benefit of participating is that other coalition members will contribute more. In practice, however, many treaties mandate a specific contribution level is independent of the number of members.

Indeed a large number of existing IEAs (a) are initially negotiated by a (smaller) set of countries which subsequently grows as additional members join and ratify the agreement over time and (b) specify pledges or duties of signatory countries that do not expand (significantly) as the number of signatories grows. Table A1 in the Appendix lists major international agreements that have these features. One example is the *Ramsar Convention on Wetlands*, the oldest of the modern global environmental agreements, which came into force in 1975 with 23 signatory countries, and has since grown to 172 contracting parties. Under this convention,

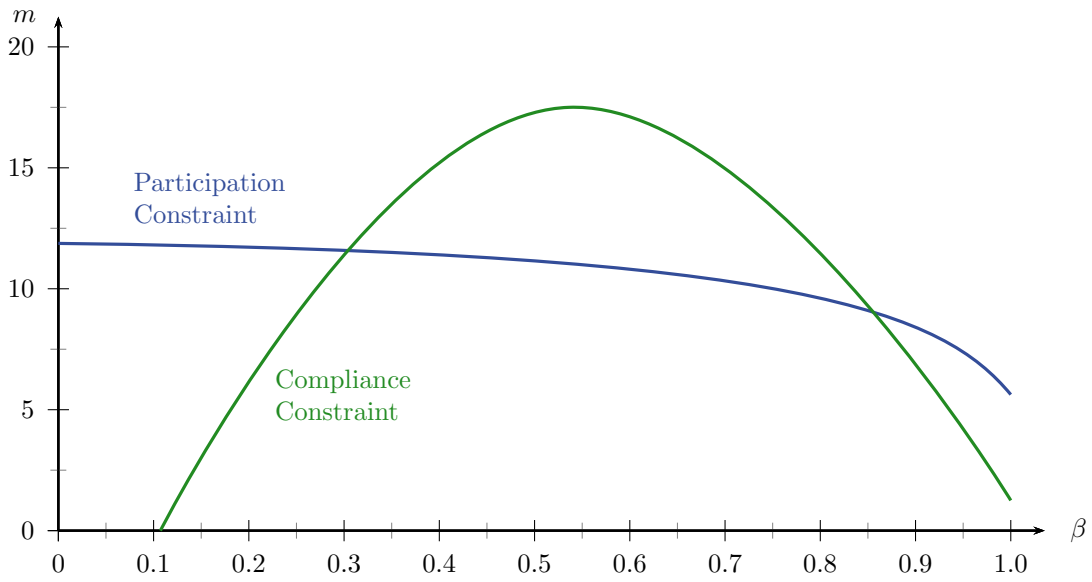

Figure 1 – Participation and Compliance. Countries will have an incentive to voluntarily enter an agreement for parameter values below the blue curve (Proposition 1). Parameter values below the green curve ensure that the agreement in self-enforcing (Proposition 2).

new countries can join only if they designate at least one wetland on their own territory as a "Ramsar" site to be included in the convention's list of protected wetlands.[15] Another example is the *Inter-American Convention for the Protection and Conservation of Sea Turtles*, which came into effect in 2001 with 11 signatory countries and currently has 16 members. Under the accord, signatory countries commit to protect or restore habitat and to take action to prevent the capture of turtles or commerce with their eggs.[16]

The traditional IEA literature on treaty size and coalition formation cannot explain these types of agreements, because the only benefit from joining a treaty in a coalition formation game is that all members of the coalition will take the spillovers on the new signatory country into account, and everyone has to increase their pledges as a result. When the members' pledges do not increase with the number of signatories, no country would wish to join a treaty. With time-inconsistent preferences, however, domestic policymakers have a unilateral incentive to join, even without the benefit from elevated commitments of existing treaty members. Moreover, we would expect this incentive to be particularly prominent if the treaty obligations pertain to an environmental action that also has a significant local component, such as protection of habitat under national jurisdiction as in the Ramsar Convention. Indeed, many existing conventions have this property (see Table A1 for details).

To fix ideas, consider the following simple game. At $t = 0$, a neutral party or "arbitrator" proposes an agreement that specifies a given increase $\alpha > 0$ in the abatement level, relative to BAU, that every member country must commit to forever after joining the treaty. Countries can sign at any time $t = 0, 1, \ldots$ and must set $a_\tau = (1 + \alpha)a^{bau}$ at every future time $\tau > t$. If m^* countries join in an MPE of the game, we will

[15] Article 2.1 of the Convention. Although the convention offers support and resources, a key commitment of the contracting parties is to manage their own Ramsar sites. With over 2,400 Ramsar sites, the convention established the world's largest network of protected areas, covering more than 2.5 million square kilometers. See https://www.ramsar.org/ for details.

[16] See http://www.iacseaturtle.org/.

call the resulting agreement a^* among m^* countries a *convention*, to distinguish it from the traditional IEA coalition- framework. The key characteristic of a convention is that domestic abatement pledges that do not vary with the number of signature countries.

Because policies under the agreement do not depend on how many countries join and because all countries have identical objective functions (up to constants) with respect to abatement levels, either all countries join, or none. For the same reason as before, every non-signatory country will set $a_{i,t} = a^{bau} \forall t$.

Proposition 4 *Suppose that a convention mandates a fixed contribution level* $a_{i,t} = (1+\alpha)a^{bau}$.

(i) If $\beta = 1$, $m^* = 0$ *for every* $\alpha > 0$.

(ii) When $\beta < 1$, *an equilibrium exists in which* $m^* = n$ *for any*

$$\alpha \leq \alpha^* \equiv 2(1-\beta)/\beta \tag{9}$$

and $m^* = 0$ *otherwise. Moreover, compliance is self-enforcing if*

$$\alpha \leq \alpha^{**} \equiv 2\frac{\delta(1-\beta)}{1-\delta(1-\beta)} < \alpha^*.$$

(iii) The convention with first-best contributions satisfies the participation constraints iff:

$$\frac{\epsilon}{\gamma} \leq \frac{1-\beta}{n-1} \tag{10}$$

and the compliance constraints iff:

$$\frac{\epsilon}{\gamma} \leq \frac{\delta(1+\beta)-1}{1-\delta(1-\beta)}\left(\frac{1-\beta}{n-1}\right), \tag{11}$$

where the r.h.s. decreases in β *at* $\beta = 1$ *iff* $\delta > 1/2$.

Part (i) confirms that if policymakers are time consistent, there is no benefit from joining any convention of this type.

Part (ii) formalizes the intuition that if governments' preferences exhibit present bias, signing a convention helps to bind future policymakers. A country's net benefit from the convention is strictly positive and maximized at $\alpha = (1-\beta)/\beta$, the value that would equate $(1+\alpha)a^{bau}$ to the current policymaker's most preferred policy (7). An arbitrator can leverage this fact and propose even higher α as long as the net benefit does not go to zero, which happens at α^*. Note that if β is smaller, a^* is larger. A decline in β from 1 increases $(1+\alpha^*)\,a^{bau}$ also if just $\delta > 1/2$.

Although it is beneficial to bind future decisions, if $\alpha > \alpha^*$ it can be tempting to defect and withdraw before complying. This is the case unless $\alpha \leq \alpha^{**}$, where we have assumed that if a country defects, it is not allowed to enter the coalition any time later.[17]

[17] By definition, conventions do not change their mandated commitments as countries join or leave. The only feasible pun-

A higher value of α raises the sum of payoffs as long as $(1+\alpha)a^{bau} < a^{FB}$. Thus, a benevolent arbitrator would propose $\alpha = \alpha^*$ (respectively, $\alpha = \alpha^{**}$ if compliance is an issue), unless the FB can be achieved. Part (iii) says that if (10) holds, the local incentive to commit is sufficiently pronounced (β and/or ϵ/γ sufficiently low) that the FB can be implemented by mandating a^{FB}; in this case, the convention constitutes a deep agreement specifying commitments that maximize the welfare of all countries. If (10) does not hold, the convention must be relatively shallow to motivate participation. Countries are tempted to defect from the first-best convention unless (11) holds. This compliance constraint is more likely to hold if β is reduced from 1, as long as $\delta > 1/2$. The intuition for this result is the same as before.

Note that the coalition size is invariant in the parameters. Because commitments do not change with m, either all or no countries will participate. There is no trade-off between broad-but-shallow versus narrow-but-deep conventions, at least not when countries are identical. If countries were heterogeneous, then weakening the mandate (by lowering α) would increase participation by inducing some non-signatory countries who previously were at the margin to adopt the convention.

While providing a rationale for the kind of agreements with fixed mandates we observe in practice, the result in Proposition 3 can provide a further insight when countries are heterogeneous. Specifically, suppose countries differed in the degree to which domestic politics suffered from time-inconsistency (β). After all, political systems differ and equilibrium political turnover (incumbency advantage) is higher in some nations than others. Since the benefit of joining a given convention is elevated for governments who face a more pronounced time-inconsistency problem, ceteris paribus, signatory countries may be those with the lowest β in equilibrium. This prediction is consistent with the fact that democracies are more likely to commit to international environmental treaties relative to autocracies, both in the cross-section and as they undergo a transition to a more democratic system.[18]

Lastly, this theory also implies that if a country defects on its pledges, or withdraws from the treaty, other signatories will not alter their contributions in conventions. All other treaty members are still motivated to participate and comply in order to deal with their own domestic time-inconsistency problem. The traditional theory, in contrast, predicts that such a withdrawal would lead to reduced participation or contributions from the other countries because, without those types of reactions, there would be no hope of inducing countries to participate and comply in the first place, given the standard assumption that they have time-consistent preferences.

In practice, this finding implies that if the US, for example, withdraws from the Paris Agreement, the other countries may not want to change their decisions regarding participation or contributions. The Paris

ishment for a defecting country therefore is to not allow future membership. Note that if $P_{i,t}$ defects, then $P_{i,t+1}$ has an incentive to enter again to commit future contributions. When $P_{i,t+1}$, in effect, renegotiates i's own strategy, without the need to persuade anyone else, $P_{i,t+1}$ will itself capture the entire surplus of skipping the punishment (Asheim, 1997). When the punishment is skipped, defection is costless, and the compliance constraint may not be satisfied. Future research should combine renegotiation concerns with the analysis above.

[18] See Battaglini and Harstad (2020, Table I). There is a large literature on international relations connecting political regimes and international cooperation. See for example Neumeyer (2002) for a cross-country analysis of IEAs. Mansfield and Pevehouse (2008) show that the likelihood of joining international environmental organizations is particularly high during the process of democratization, and speculate that a desire to commit (to reforms and future policies) can explain their finding. It is straightforward to allow for heterogeneity when the abatement mandate is fixed, because then participation depends only on domestic parameter values. If, instead, the abatement level is decided on collectively, as in Section 3.1, heterogeneity can influence the level of ambition and generate additional time-inconsistency problems (Bowen et al., 2019).

Agreement has features similar to that of a convention because the nationally determined contributions are not contingent on what other countries do change with the withdrawal of a treaty party. Of course, actual international environmental agreements tend to share some characteristics with conventions, while other characteristics are better captured in the traditional model of treaty negotiations; hence, each formulation touches on parts of the truth. The argument here is that, while the traditional theory requires readjustments of contributions or changes in participation of others to motivate a country to comply and participate, our theory can explain participation and compliance without the need to predict such severe consequences in case of an exit by a treaty member, especially if the country is as significant as the U.S. In this sense, our model points to a more optimistic view about the future for climate cooperation.

4 Endogenous Elections

So far, the probability of staying in power has been exogenous. In fact, the equilibrium re-election probability p must be constant in the model above when voters use Markov strategies because candidates are identical and they cannot commit to future platforms. Thus, past actions are not relevant for the future, and Markov strategies at the voting stage will not be contingent on them.

If voters are backward looking, they may be inclined to reward an incumbent that has delivered a desirable contribution.[19] The utility that $P_{i,t}$ delivers in BAU can be measured by:

$$U_{i,t}\left(g_{i,t}\right) = -\frac{b}{2}a_{i,t}^2 + \delta\gamma C a_{i,t}.$$

Using a simple probabilistic voting framework, i's probability of remaining in power is increasing in the utility that is offered and can be written as (e.g., Persson and Tabellini, 2000)

$$p + \left[U_{i,t} - U_{i,t}^{Eq.}\right]\phi,$$

where p and ϕ are positive constants and $U_{i,t}^{Eq.}$ is the equilibrium level of $U_{i,t}$. If R is the exogenously given "office rent", $P_{i,t}$'s benefit from winning the next election is $\delta\left[\Delta u^{Eq.} + R\right]$, assumed to be positive, where $u_{i,t+1} = u^{Eq.}$ is given by the equilibrium decisions at $t+1$. With these modifications, $P_{i,t}$'s preferred $a_{i,t}$ will solve:

$$\max_{a_{i,t}} \left(1+\Delta\right)u_{i,t} + \left[U_{i,t} - U_{i,t}^{Eq.}\right]\phi\delta\left[\Delta u^{Eq.} + R\right] + \sum_{\tau=t+1}\delta^{\tau-t}\left[\left(1+p\Delta\right)u^{Eq.} + pR\right]$$

$$\Rightarrow \qquad a_{i,t} = a_{\phi}^{bau} \equiv \beta_{\phi}\frac{\delta\gamma C}{b}, \text{ where } \beta_{\phi} \equiv \frac{1+p\Delta+\phi\delta\left(\Delta u^{Eq.}+R\right)}{1+\Delta+\phi\delta\left(\Delta u^{Eq.}+R\right)}.$$

Consequently, $P_{i,t}$ continues to face a time inconsistent problem. As before, $P_{i,t}$ would like to commit to $a_{\tau} = \delta\gamma C/b$, $\tau > t$, but in equilibrium, $a_{i,\tau} = \beta_{\phi}\delta\gamma C/b$. If ϕ or the office rent is large, $P_{i,t}$ pays more attention to $U_{i,t}$, and β_{ϕ} will be larger. Thus, backward-looking voters can weaken the time inconsistency

[19] Prominent theories of electoral competition where voters are backward looking include for example the retrospective voting model (e.g. Ferejohn, 1986) as well career-concerns or other models where an officeholder's actions or performance in the previous period carries information on future actions or performance (see, e.g., Persson and Tabellini (2000).

problem in BAU, but it is not eliminated: For every finite ϕ and R, $\beta_\phi \in (0,1)$ if $p\Delta \in (0,\Delta)$.

If voters reward policymakers that sign treaties, because the treaty will raise the voters' future utility, then the motivation to sign an IEA is strengthened relative to the analysis in Section 3. If candidates are heterogeneous, however, it is not clear that voters will act in this way. Because forward-looking voters will elect the candidate that will provide more utility in the next period, the incumbent may attempt to differentiate herself from the challengers. This attempt will influence whether or not she will benefit from an IEA. (For an analysis of this situation, see Battaglini and Harstad, 2020).

5 Concluding Remarks

This paper sheds light on how international treaties can draw on domestic time-inconsistency problems. In our framework, present bias arises because policymakers rotate being in office. Thus, each government would like future governments to act sustainably but, once in office, is tempted to postpone costly actions. We show that the larger the domestic time-inconsistency problem, the greater the incentive to tie the hands of future policymakers, and the larger the equilibrium coalition size of IEAs. Further, the motivation to comply with an agreement, rather than to defect, can be stronger if domestic policy preferences exhibit a present bias. The positive effect of present bias on participation and compliance is more pronounced when the international spillovers are limited relative to the domestic policy issue.

Present bias in domestic politics can also motivate countries to sign international environmental agreements specifying pledges that are not renegotiated as additional signatories join. The traditional theory on IEAs cannot explain such arrangements, because the only benefit of joining a treaty in the standard model is that participation leads to higher contribution levels by all the other members. Yet, this type of agreement is frequently observed in practice in areas such as habitat preservation or species protection, and even the Paris Agreement is of this type. If the US exits the Paris Agreement, our theory predicts that other countries may remain and contribute as before, because their pledges are made to commit future domestic policymakers. The traditional theory, in contrast, would predict that other countries would also exit or reduce their own contributions following the exit of the US.

Our argument was based on a simple model that abstracted from elections, heterogeneity, and asymmetric information, and other relevant factors.

In our framework, international agreements are the sole instrument to bind future governments. Naturally, if domestic policymakers had other means to tie the hands of their successors, the benefit of international treaties would be diminished, *ceteris paribus*. Indeed, as would be the case with time consistent preferences, the maximum feasible coalition only contains three countries and conventions would not form if we allowed policymakers to fully commit to future actions by other means.[20] This argument highlights that domestic time-inconsistency problems could be key to explain both the large size and the type of coalitions we observe in reality.

It is important to learn more about the extent to which policymakers adopt strategies that yield consistent

[20] See the Appendix A for a formal argument.

policies, with and without international treaties. Theoretical and empirical research along these lines is necessary to deepen our understanding of how international agreements should be designed so that they do not simply account for domestic political failures, but rather take advantage of them to facilitate participation and compliance.

Appendix A: The Role of Commitment

To formally highlight the critical role of commitment that international treaties play, suppose that $P_{i,t}$ at time t can fully commit to future policies a_τ for every $\tau > t$ a treaty. We have

Corollary *Suppose $P_{i,t}$ can commit to future policies without the treaty.*

(i) A deep agreement can be of size m if and only if $m \leq 3$.

(ii) For a convention with mandate $\alpha > 0$, in every equilibrium, $m = 0$.

Part (i) follows trivially from Proposition 1 since countries no longer derive the additional benefit of self-commitment. Part (ii) reiterates our earlier claim that in the absence of a domestic time-inconsistent policy problem, no one would participate in a convention when the other signatories face the same mandate regardless of the number of members.

Table A.1 – List of Major International Conventions

Treaty Name (year in effect)	Signatories	Parties to Date	Primary Policy Issue	Locally Relevant Commitment	Source
Aarhus Convention on citizen's rights regarding governmental decision-making processes on environmental matters (2001)	38 (incl EU)	47 (incl EU)	environment as a human rights issue, increases government accountability, transparency and responsiveness	convention explicitly covers rights regarding policies affecting local and national environment environment	https://unece.org/environment-policy/public-participation/aarhus-convention/
Basel Convention on Movements of Hazardous Wastes and their Disposal (1992)	53	191	protecting vulnerable countries from unwanted hazardous waste imports	national waste management policy and strategy	https://www.basel.int/
Convention on the Protection of Migratory Species of Wild Animals (1983)	13	133	conservation and sustainable use of migratory animals and their habitats	maintaining a local network of habitats, providing new local habitats and reintroduce migratory species into suitable local habitats	https://minamataconvention.org/en
Minamata Convention on Mercury (2013)	128	152	regulations on use of, and pollution from, mercury use and ban of mercury mines	closing domestic mines, phasing out of use of mercury in production, and local pollution control	https://minamataconvention.org/en
Ramsar Convention on Wetlands (1975)	23	172	managing and protecting wetlands	at least one one domestic wetland is designated as 'Ramsar site' and receives protection	https://www.ramsar.org
Stockholm Convention on Persistent Organic Pollutants (2001)	152	186	protection from chemicals that remain intact for long periods	prohibit or eliminate domestic production and use of persistent organic pollutants	https://chm.pops.int/
UN Convention on Combat of Desertification (1996)	114	194	Addressing desertification and the effects of drought in affected countries	new legislation and developing new strategies to address local desertification	https://www.unccd.int

Notes. The term 'signatory' denotes states that were engaged in the treaty process and signed onto the original treaty text. Typically, this initial show of support is followed by ratification within a few years of signing. The term 'party' refers to states that had ratified the agreement by February 2025, thereby explicitly consenting to be bound by the treaty. All of the listed conventions monitor compliance and the majority has formal bodies such as compliance committees with legal powers to help enforcement.
The sources of the information displayed in the table are the respective treaty web-pages (see Table, all accessed 02/20/2025) as well as the UN Treaty Collection https://treaties.un.org/ (accessed 2/19/2025).

Appendix B: Proofs

Proof of Proposition 1.

For every $i \in M$, it is easy to check that the payoff, relative to BAU, is a function of the $a_{j,\tau}$'s that is independent of the bliss points, i.e., the $g^*_{j,\tau}$'s. Thus, when the bargaining solution predicts an efficient and symmetric outcome when the bargaining set is symmetric (as does the Nash Bargaining Solution, for example), then, at every future time $\tau > t$, $a_{i,\tau}$ is:

$$a(m) = \delta(\gamma + (m-1)\epsilon) C/b,$$

which takes into account the externality on $m-1$ other coalition members. With this, $P_{i,t}$'s continuation value after signing is $(1+p\Delta)$ multiplied by:

$$v^{in}_{i,t}(m) = \sum_{\tau=t+1}^{\infty} \delta^{\tau-t} \Bigg[-\frac{b}{2} a(m)^2 - \delta\gamma C \left(g^*_{i,\tau} - a(m)\right) - \delta\epsilon C \sum_{j \in N\setminus i} g^*_{j,\tau} + \delta\epsilon C (m-1) a(m) + \delta\epsilon C (n-m) a^{bau} \Bigg].$$

Conversely, the continuation value if $P_{i,t}$ had chosen to free ride instead would have been $(1+p\Delta)$ multiplied by:

$$v^{out}_{i,t}(m-1) = \sum_{\tau=t+1}^{\infty} \delta^{\tau-t} \Bigg[-\frac{b}{2} \left(a^{bau}\right)^2 - \delta\gamma C \left(g^*_{i,\tau} - a^{bau}\right) - \delta\epsilon C \sum_{j \in N\setminus i} g^*_{j,\tau} + \delta\epsilon C (m-1) a(m-1) + \delta\epsilon C (n-m) a^{bau} \Bigg].$$

For internal stability, $v^{in}_{i,t}(m) \geq v^{out}_{i,t}(m-1)$, implying:

$$\begin{aligned}
-\frac{b}{2}\left(\frac{\delta C(\gamma + (m-1)\epsilon)}{b}\right)^2 + \frac{b}{2}\left(\beta\frac{\delta C\gamma}{b}\right)^2 + \delta\gamma C\left(\frac{\delta C(\gamma(1-\beta) + (m-1)\epsilon)}{b}\right) + \frac{(\delta C\epsilon)^2}{b}(m-1) \geq 0 \Leftrightarrow \\
-\frac{1}{2}\left[(\gamma + (m-1)\epsilon)^2 - (\beta\gamma)^2\right] + \gamma\left[\gamma(1-\beta) + (m-1)\epsilon\right] + (m-1)\epsilon^2 \geq 0 \Leftrightarrow \\
\frac{1}{2}\left[\Upsilon^2 - (\beta\gamma)^2\right] - \gamma(\Upsilon - \beta\gamma) - \Upsilon\epsilon + \gamma\epsilon \leq 0 \Leftrightarrow \\
\frac{1}{2}\Upsilon^2 - (\gamma+\epsilon)\Upsilon - \frac{1}{2}\left[(\beta\gamma)^2 - 2\beta\gamma^2 - 2\gamma\epsilon\right] \leq 0,
\end{aligned}$$

if we define $\Upsilon := \gamma + (m-1)\,\epsilon$. Thus, $P_{i,t}$ is indifferent between joining the treaty and not joining if

$$\begin{aligned}
\Upsilon &= \gamma + \epsilon + \sqrt{(\gamma+\epsilon)^2 + (\beta\gamma)^2 - 2\beta\gamma^2 - 2\gamma\epsilon} \Leftrightarrow \\
\gamma + (m-1)\,\epsilon &= \gamma + \epsilon + \sqrt{(\gamma+\epsilon)^2 + (\beta\gamma)^2 - 2\beta\gamma^2 - 2\gamma\epsilon} \Leftrightarrow \\
m &= 2 + \frac{1}{\epsilon}\sqrt{(\gamma+\epsilon)^2 - (\beta\gamma)^2 - 2\gamma\epsilon} = 2 + \sqrt{\left(\frac{\gamma}{\epsilon}+1\right)^2 - \left(\frac{\beta\gamma}{\epsilon}\right)^2 - 2\frac{\gamma}{\epsilon}} \Leftrightarrow \\
m &= \widehat{m} \equiv 2 + \sqrt{1 + \left(\frac{\gamma}{\epsilon}\right)^2 (1-\beta^2)} = 2 + \sqrt{1 + \frac{1-\beta^2}{(\epsilon/\gamma)^2}}.
\end{aligned}$$

For coalition sizes $m \leq m^*$, a member benefits from participating, while, if $m > m^*$, a member would strictly benefit from not participating. □

Proof of Proposition 2.

(i) By defecting on some pledged a, i will in this period, and forever after, change to $a^{bau} = \beta\delta\gamma C/b$. Holding fixed the other $a_{j,t}$'s, this benefit can be measured by:

$$\left(1 + \frac{\delta\beta}{1-\delta}\right)\frac{b}{2}\left(a^2 - \left(a^{bau}\right)^2\right) - \frac{\delta\beta}{1-\delta}\gamma C\left(a - a^{bau}\right).$$

If defection leads all other treaty members, one period later, to change their abatements to a^{bau}, with probability q, the cost of defecting is:

$$q\,(m-1)\,\frac{\beta\delta^2}{1-\delta}\epsilon C\left(a - a^{bau}\right).$$

Combined, defecting is unattractive if:

$$\begin{aligned}
\left(1 + \frac{\delta\beta}{1-\delta}\right)\frac{b}{2}\left(a^2 - \left(a^{bau}\right)^2\right) &\leq \frac{\beta\delta}{1-\delta}\gamma C\left(a - a^{bau}\right) + q\,(m-1)\,\frac{\beta\delta^2}{1-\delta}\epsilon C\left(a - a^{bau}\right) \Leftrightarrow \\
\left(1 + \frac{\delta\beta}{1-\delta}\right)\frac{b}{2}\left(a + a^{bau}\right) &\leq \frac{\beta\delta}{1-\delta}\gamma C + q\,(m-1)\,\frac{\beta\delta^2}{1-\delta}\epsilon C \Leftrightarrow \\
(1-\delta+\delta\beta)\,(ab/\beta + \delta\gamma C) &\leq 2\delta\gamma C + 2q\,(m-1)\,\delta^2\epsilon C \Leftrightarrow \\
(1-\delta+\delta\beta)\,(ab/\beta\delta) + \delta\beta\gamma C &\leq \gamma C\,(1+\delta) + 2q\,(m-1)\,\delta\epsilon C \Leftrightarrow \qquad (12) \\
a &\leq \frac{\gamma\,(1+\delta\,(1-\beta)) + 2q\,(m-1)\,\delta\epsilon}{1-\delta\,(1-\beta)}\frac{\beta\delta C}{b} \text{ or} \qquad (13) \\
a &\leq (1+\alpha)\,a^{bau}, \text{ where } \alpha := \frac{2\delta\,(1-\beta) + 2q\,(m-1)\,\delta\epsilon/\gamma}{1-\delta\,(1-\beta)}, \qquad (14)
\end{aligned}$$

where α decreases in β. The best SPE is symmetric, so $m = n$.

(ii) The first-best a^{FB} satisfies (13) iff:

$$\begin{aligned} \delta C \frac{\gamma + (n-1)\epsilon}{b} &\leq \frac{\gamma(1+\delta(1-\beta)) + 2q(n-1)\delta\epsilon}{1-\delta(1-\beta)} \frac{\beta\delta C}{b} \Leftrightarrow \\ [\gamma + (n-1)\epsilon][1-\delta(1-\beta)] &\leq [\gamma(1+\delta(1-\beta)) + 2q(n-1)\delta\epsilon]\beta, \end{aligned}$$

which gives (ii) after some algebra. (iii) The derivative of the l.h.s. of (12) w.r.t. β is positive iff:

$$\frac{\beta^2\delta}{1-\delta} > \frac{ab}{\delta\gamma C}. \tag{15}$$

At (13), this implies:

$$\begin{aligned} \frac{\beta^2\delta}{1-\delta} &> \frac{b}{\delta\gamma C} \frac{\beta[\gamma(1+\delta(1-\beta)) + 2q(m-1)\delta\epsilon]\delta C/b}{1-\delta(1-\beta)} \Leftrightarrow \\ \frac{\beta\delta}{1-\delta} &> \frac{1+\delta(1-\beta) + 2q(m-1)\delta\epsilon/\gamma}{1-\delta(1-\beta)} \Leftrightarrow \\ \frac{\beta\delta}{1-\delta} &> \frac{2+2q(m-1)\delta\epsilon/\gamma}{1-\delta(1-\beta)} - 1 \Leftrightarrow \\ \frac{1-\delta(1-\beta)}{1-\delta} &> \frac{2+2q(m-1)\delta\epsilon/\gamma}{1-\delta(1-\beta)} \Leftrightarrow \\ \frac{[1-\delta(1-\beta)]^2/2}{1-\delta} &> 1 + q(m-1)\delta\epsilon/\gamma. \end{aligned}$$

So, when this holds, a larger β requires a to decline when (12) binds. □

Proof of Proposition 3.

(i) With $a = a(m)$, (13) becomes:

$$\begin{aligned} (\gamma + (m-1)\epsilon)\frac{\delta C}{b} &\leq \frac{\gamma(1+\delta(1-\beta)) + 2q(m-1)\delta\epsilon}{1-\delta(1-\beta)} \frac{\beta\delta C}{b} \Leftrightarrow \\ (\gamma + (m-1)\epsilon)[1-\delta(1-\beta)] &\leq [\gamma(1+\delta(1-\beta)) + 2q(m-1)\delta\epsilon]\beta \Leftrightarrow \\ (m-1)[1-\delta(1-\beta) - 2q\delta\beta]\epsilon/\gamma &\leq [1+\delta(1-\beta)]\beta - [1-\delta(1-\beta)] \Leftrightarrow \\ [1-\delta(1-\beta(1-2q))]\epsilon/\gamma &\leq [\delta(1+\beta) - 1](1-\beta)/(m-1). \end{aligned}$$

If $\beta = 1$, the inequality boils down to $\delta q \geq 1/2$. If, instead, $q = 1$, the inequality implies $\delta(1+\beta) \geq 1$. More generally, the inequality always fails if $\delta(1+\beta) < 1$. But if the inequality binds, both brackets must be positive, so it implies:

$$\frac{\epsilon}{\gamma}(m-1) \leq \frac{(\delta[1+\beta] - 1)(1-\beta)}{1-\delta(1-\beta[1-2q])} \tag{16}$$

(ii) The derivative of the r.h.s. w.r.t. β is:

$$\begin{aligned} & [\delta(1-\beta) - \delta[1+\beta] + 1][1-\delta[1-\beta(1-2q)]] - \delta(1-2q)(\delta[1+\beta] - 1)(1-\beta) \\ = \; & -[2\delta\beta - 1][1-\delta[1-\beta(1-2q)]] - \delta(1-2q)(\delta[1+\beta] - 1)(1-\beta). \end{aligned}$$

When $\beta \to 1$, this becomes $1-2\delta$. Thus, when $\delta > 1/2$, a decline in β from 1 expands the set of parameters for ϵ and γ under which the compliance constraint holds. □

Proof of Proposition 4.

Consider a proposed increase $\alpha > 0$ over a^{bau}, so that if $P_{i,t}$ signs the convention then country i commits to $a = (1+\alpha)a^{bau}$ for every period $\tau > t$. We can write i's net benefit from signing a convention with $m-1$ other signatories, relative to opting out, as $(1+p\Delta)$ multiplied by

$$\begin{aligned} \upsilon(\alpha) &= \sum_{\tau=t+1}^{\infty} \delta^{\tau-t}\left[-\frac{b}{2}\left((1+\alpha)a^{bau}\right)^2 + \frac{b}{2}\left(a^{bau}\right)^2 + \delta\gamma C\alpha a^{bau}\right] \\ &= \frac{\delta}{1-\delta}\alpha a^{bau}\left[-b\left(1+\frac{1}{2}\alpha\right)a^{bau} + \delta\gamma C\right], \end{aligned}$$

which is independent of m (as expected). For any $\alpha > 0$, $\upsilon(\alpha) \geq 0$ if and only if

$$b(1+\alpha/2)\,a^{bau} = b(1+\alpha/2)\,\beta\delta\gamma C/b \leq \delta\gamma C \Leftrightarrow \alpha \leq \alpha^* \equiv 2(1/\beta - 1).$$

(i) Hence, $\beta = 1$ implies $\upsilon(\alpha) < 0$ for any $\alpha > 0$.

(ii) For $\beta < 1$, $\upsilon(\alpha) > 0$ for $\alpha \in (0, \alpha^*)$ and in the unique MPE, every $P_{i,t}$ will sign such a convention. At $\alpha = \alpha^*$, $P_{i,t}$ is indifferent between joining and not joining, so there is an MPE in which everyone participates.

Suppose that if a country defects, it will not enter the coalition any time later. When this is the only consequence, the convention is self-enforcing iff:

$$\begin{aligned} b\left(1+\frac{1}{2}\alpha\right)a^{bau}\left(1+\frac{\beta\delta}{1-\delta}\right) &\leq \delta\gamma C\left(\beta+\frac{\beta\delta}{1-\delta}\right) \Leftrightarrow \\ \left(1+\frac{1}{2}\alpha\right)\left(1+\frac{\beta\delta}{1-\delta}\right) &\leq \left(1+\frac{\delta}{1-\delta}\right) \Leftrightarrow \\ \left(1+\frac{1}{2}\alpha\right)(1-\delta(1-\beta)) &\leq 1 \Leftrightarrow \\ \frac{1}{2}\alpha(1-\delta(1-\beta)) &\leq \delta(1-\beta) \Leftrightarrow \\ \alpha \leq \alpha^{**} := 2\frac{\delta(1-\beta)}{1-\delta(1-\beta)} &< \alpha^*. \end{aligned}$$

(iii) An arbitrator who seeks to maximize payoffs or aggregate abatement levels would set $\alpha = \alpha^*$, unless the first best is implementable. The FB abatement levels satisfies the participation constraints for the members

iff:

$$\begin{aligned}
(1+\alpha^{*})a^{bau} &\geq a^{FB} \Leftrightarrow \\
\left(1+2\frac{1-\beta}{\beta}\right)\beta\gamma &\geq \gamma+(n-1)\epsilon \Leftrightarrow \\
2-\beta &\geq 1+(n-1)\epsilon/\gamma \Leftrightarrow \\
\beta &\leq 1-(n-1)\epsilon/\gamma.
\end{aligned}$$

The FB is also self-enforcing as a convention iff the compliance constraint is satisfied:

$$\begin{aligned}
(1+\alpha^{**})a^{bau} &\geq a^{FB} \Leftrightarrow \\
\left(1+2\frac{\delta(1-\beta)}{1-\delta(1-\beta)}\right)\beta\gamma &\geq \gamma+(n-1)\epsilon \Leftrightarrow \\
\left(\frac{2\delta\beta}{1-\delta(1-\beta)}-1\right)(1-\beta) &\geq (n-1)\epsilon/\gamma \Leftrightarrow \\
\left(\frac{2\delta\beta-1+\delta-\delta\beta}{1-\delta(1-\beta)}\right)(1-\beta) &\geq (n-1)\epsilon/\gamma \Leftrightarrow \\
(n-1)\epsilon/\gamma &\leq \frac{\delta(1+\beta)-1}{1-\delta(1-\beta)}(1-\beta)
\end{aligned}$$

Note that the derivative of the r.h.s. w.r.t. β is the same as for (16) at $q=0$. Consequently, a decline in β from 1 expands the set of parameters for ϵ and γ under which the compliance constraint holds iff $\delta > 1/2$. □

References

Aldy, Joseph E., and Robert N. Stavins. 2009. *Post Kyoto International Climate Policy: Summary for Policymakers*. New York: Cambridge University Press.

Alesina, Alberto, and Guido Tabellini. 1990. "A Positive Theory of Fiscal Deficits and Government Debt in a Democracy." *Review of Economic Studies* 57: 403–14.

Amador, Manuel. 2003. "A Political Economy Model of Sovereign Debt Repayment." *mimeo*, Stanford Graduate School of Business.

Asheim, Geir B. 1997. "Individual and Collective Time-Consistency." *The Review of Economic Studies* 64(3): 427–43.

Barrett, Scott. 1994, "Self-Enforcing International Environmental Agreements." *Oxford Economic Papers* 46: 878–94.

Barrett, Scott. 2005. "The Theory of International Environmental Agreements." In *Handbook of Environmental Economics (Vol 3)*, edited by K. G. Mäler and J. R. Vincent, 1457–1516. Elsevier.

Battaglini, Marco, and Bård Harstad. 2016."Participation and Duration of Environmental Agreements." *Journal of Political Economy* 124(1): 160–204.

Battaglini, Marco, and Bård Harstad. 2020. "The Political Economy of Weak Treaties." *Journal of Political Economy* 128(2): 544–90.

Bisin, Alberto, Alessandro Lizzeri, and Leeat Yariv. 2015. "Government Policy with Time Inconsistent Voters." *American Economic Review* 105(6): 1711–37.

Carraro, Carlo, and Domenico Siniscalco. 1993. "Strategies for the International Protection of the Environment." *Journal of Public Economics* 52(3): 309–28.

Chatterjee, Satyajit, and Burcu Eyigungor. 2016. "Continuous Markov Equilibria with Quasi-Geometric Discounting." *Journal of Economic Theory* 163: 467–94.

Coconi, Paula, and Carlo Perroni. 2009. "Do Credible Domestic Institutions Promote Credible International Agreements?" *Journal of International Economics* 79: 160–70.

DellaVigna, Stefano, and Ulrike Malmendier. 2006. "Paying Not to Go to the Gym." *American Economic Review* 96(3): 694–719.

Ferejohn, John. 1986. "Incumbent Performance and Electoral Control." *Public Choice* 50: 5–26.

Finus, Muchael, and Stefan Maus. 2008. "Modesty May Pay." *Journal of Public Economic Theory* 10: 801–26.

Fischer, Stanley. 1980. "Dynamic Inconsistency, Cooperation, and the Benevolent Dissembling of Government." *Journal of Economic Dynamics and Control* 2: 93–103.

Gollier, Christian, and Martin L. Weitzman. 2010. "How Should the Distant Future Be Discounted When Discount Rates Are Uncertain?" *Economics Letters* 107(3): 350–53.

Gollier, Christian, and Richard Zeckhauser. 2005. "Aggregation of Heterogeneous Time Preferences." *Journal of Political Economy* 113(4): 878-96.

Harstad, Bård. 2023a. "Pledge-and-Review Bargaining: From Kyoto to Paris." *The Economic Journal* 133 (651): 1181–1216.

Harstad, Bård. 2023b. "The Conservation Multiplier." *Journal of Political Economy* 131(7): 1731–71.

Harstad, Bård. 2020. "Technology and Time Inconsistency." *Journal of Political Economy* 128(7): 2653–89.

Hoel, Michael. 1992. "International Environment Conventions: The Case of Uniform Reductions of Emissions." *Environmental and Resource Economics* 2: 141–59.

Karp, Larry, and Leo Simon. 2013. "Participation Games and International Environmental Agreements: A Non-parametric Model." *Journal of Environmental Economics and Management* 65(2): 326-344.

Kehoe, Patrick J. 1989. "Policy Cooperation among Benevolent Governments May Be Undesirable." *Review of Economic Studies* 56: 289–96.

Kydland, Finn E., and Edward C. Prescott. 1977. "Rules Rather than Discretion: The Inconsistency of Optimal Plans." *Journal of Political Economy* 85(3): 473–91.

Laffont, Jean-Jacques, and Jean Tirole. 1996. "Pollution Permits and Compliance Strategies." *Journal of Public Economics* 62: 85–125.

Laibson, David. 1997. "Golden Eggs and Hyperbolic Discounting." *Quarterly Journal of Economics* 112(2): 443–78.

Maggi, Giovanni, and Andres Rodriguez-Clare. 1998. "The Value of Trade Agreements in the Presence of Political Pressures." *Journal of Political Economy* 106(3): 574–601.

Maggi, Giovanni, and Andres Rodriguez-Clare. 2007. "A Political-Economy Theory of Trade Agreements." *American Economic Review* 97(4): 1374–1406.

Mailath, George J., and Larry Samuelson. 2006. *Repeated Games and Reputations: Long Run Relationships*, Oxford: Oxford University Press.

Mansfield, Edward D., and John C. Pevehouse. 2008. "Democratization and the Varieties of International Organizations." *Journal of Conflict Resolution* 52(2): 269–94.

Matsuyama, Kiminori. 1990, "Perfect Equilibria in a Trade Liberalization Game." *American Economic Review* 80(3): 480–92.

Neumayer, Eric. 2002. "Do Democracies Exhibit Stronger International Environmental Commitment? A Cross-Country Analysis." *Journal of Peace Research* 39(2): 139–64.

Persson, Torsten, and Lars Svensson. 1989. "Why a Stubborn Conservative Would Run a Deficit: Policy with Time Inconsistency Preferences." *Quarterly Journal of Economics* 104(2): 325–45.

Persson, Torsten, and Guido Tabellini. 2000. *Political Economics: Explaining Economic Policy*. Cambridge, MA: MIT Press.

Phelps, Edmund S., and Robert A. Pollak. 1968. "On Second-Best National Saving and Game-Equilibrium Growth." *Review of Economic Studies* 35: 165–99.

Ploeg, Frederick van der. 1988. "International Policy Coordination in Interdependent Monetary Economies." *Journal of International Economics* 25(1–2): 1–23.

Rogoff, Kenneth. 1985."Can International Monetary Cooperation be Counterproductive." *Journal of International Economics* 18: 199–217.

Spycher, Sarah. 2024. "Elections and Political Polarisation: Challenges for Environmental Agreements." *Quaderni - Working Paper DSE N.1196* Department of Economics, University of Bologna.

Staiger, Robert, and Guido Tabellini. 1999. "Do GATT Rules Help Governments Make Domestic Commitments?" *Economics and Politics* 11(2): 109–44.

Tabellini, Guido. 1991. "The Politics of Intergenerational Redistribution." *Journal of Political Economy* 99(2): 335–57.

Weitzman, Martin L. (2001). "Gamma Discounting." *American Economic Review* 91(1): 260—71.