THE ECONOMIC JOURNAL

The Economic Journal, **133** (*April*), 1181–1216 https://doi.org/10.1093/ej/ueac076
Advance Access Publication Date: 9 November 2022

PLEDGE-AND-REVIEW BARGAINING: FROM KYOTO TO PARIS*

Bård Harstad

A tractable dynamic model of international climate policies is analysed. The choice of bargaining game influences participation levels, emission quotas and technology investment levels. I derive several predictions that are arguably consistent with the differences between the 1997 Kyoto Protocol and the 2015 Paris Agreement—including the transitioning from the former to the latter.

The pledge-and-review strategy is completely inadequate.

Christian Gollier and Jean Tirole
The Economist, June 1, 2015

The climate agreements signed in Kyoto, 1997, and Paris, 2015, were different in several respects. My goals in this paper are to investigate how different bargaining procedures influence climate policies and to understand the determinants of the best procedure.

Pledge and review (P&R) refers to the structure of the negotiations associated with the Paris Agreement. Before the countries were expected to sign the climate agreement, each party was asked to submit an intended nationally determined contribution (NDC). For most developed countries, the NDC specified unilateral cuts in the emissions of greenhouse gases being effective from 2020 to 2025 (or to 2030). As an illustration, Table 1 presents the pledges for a sample of developed countries.[1]

Every five years the parties shall review and make new pledges for another five-year period (Paris Agreement Art. 4.9).

This procedure is remarkably different from that used for the Kyoto Protocol. There, a 'top-down' approach was used to pressure governments to cut emissions by (on average) 5% relative to

* Corresponding author: Bård Harstad, Department of Economics, University of Oslo, Pb 1095 Blindern, 0317 Oslo, Norway. Email: bard.harstad@econ.uio.no

This paper was received on 10 June 2020 and accepted on 2 October 2022. The Editor was Estelle Cantillon.

I thank Scott Barrett, Ernesto Dal Bo, Faruk Gul, Jon Hovi, Steffen Lippert, Paolo Piacquadio, Santiago Rubio, Leo Simon, Håkon Sælen, David Victor, Christina Voigt, Joel Watson and audiences in Adelaide (AARES pre-conference), Universitat Autònoma de Barcelona, University of Barcelona, the BEET workshop at BI, UC Berkeley, UC3M, CEU, UC San Diego, University of Chicago, CREST-Ecole Polytechnique, EIEF, ESEM 2018, University of Essex, HEC Paris, Hong Kong Baptist University, Ifo Institute, London School of Economics, Manchester University, University of Melbourne, MIT, National Taiwan University, National University of Singapore, Northwestern University, CSEF-University of Naples Federico II, University of NotreDame, University of Oslo, UPF, Princeton University, Queen Mary University, the San Francisco Fed., Singapore Management University, Stanford GSB, SURED 2018, Toulouse School of Economics, the 2019 Wallis Conference and WCERE 2018. Marie Karlsen and Johannes Hveem Alsvik provided excellent research assistance and Frank Azevedo helped with the copyediting.

This research received funding from the European Research Council under the EU's 7th Framework Programme, ERC GA no. 683031.

[1] The baseline year is 1990 for the European Union, Russia and Switzerland, while it is 2005 for Australia, Canada, New Zealand and the United States. Article 4.4 of the Paris Agreement encourages 'economy-wide absolute emission reduction targets', although several developing countries state pledges in terms of emission per GDP and some of these are conditional on receiving transfers. The official list is available at http://www4.unfccc.int/ndcregistry, but, for an overview, see http://cait.wri.org/indc/#/.

Table 1. *Pledges Specifying Emission Cuts Relative to Nationally Chosen Baselines.*

Party	Australia	Canada	EU	New Zealand	Norway	Russia	USA
Pledge	26%–28%	30%	40%	30%	40%	25%–30%	26%–28%

the 1990 levels. Bodansky and Rajamani (2018, p. 23) found that: 'In essence, the Kyoto Protocol was the product of mutual concessions [and, as a consequence] USA accepted a much stronger target (minus 7% from 1990 levels) than it had wanted ....' The conditionality of concessions in Kyoto makes it unsurprising that most papers have associated the bargaining outcome with the Nash bargaining solution (NBS); see the literature review below.

By comparison, P&R has been referred to as a 'bottom-up' approach because countries themselves determine how much to cut nationally. According to the Paris Agreement (Art. 4.2): 'Each Party shall prepare, communicate and maintain successive nationally determined contributions that it intends to achieve.' Based on this, Victor (2015) observed that: 'Now, instead of setting commitments through centralized bargaining, the Paris approach sets countries free to make their own commitments' (Victor, 2015). *The New York Times* (November 28, 2015) wrote that: 'Instead of pursuing a [Kyoto-style] top-down agreement with mandated targets, [the organizers] have asked every country to submit a national plan that lays out how and by how much they plan to reduce emissions in the years ahead.'

Several observers and scholars fear that P&R is unable to deliver as ambitious targets as would a top-down approach. Keohane and Oppenheimer (2016, p. 142) predicted that: 'Many governments will be tempted to use the vagueness of the Paris Agreement, the discretion that it permits, to limit the scope or intensity of their proposed actions.' Tirole (2017, p. 209) added: 'The strategy of voluntary commitments has several significant defects, and is an inadequate response to the climate change challenge.' My companion paper (Harstad, 2023) formalises the P&R bargaining game and proves that predicted contributions are smaller than under the NBS.

This paper explores the consequences of different bargaining procedures. Alternative bargaining outcomes are embedded in a new tractable climate policy game. I find that, with P&R, relative to the NBS, the pledged emission cuts are smaller, investments in renewables are smaller and everyone's payoff is therefore lower. The negative result is reversed, however, when we account for free riding and let the decision to participate in the bargaining game be voluntary. Under the P&R procedure, a party places a lower weight on the interest of others (Harstad, 2023), and thus it is not that costly for a party to participate, and the equilibrium coalition size is larger than with the NBS. The larger coalition size means, in equilibrium, that the *sum* of contributions is larger with P&R, the aggregate investments are larger and so is welfare.

The comparison of bargaining procedures is more interesting when we take into account the fact that there is an upper boundary ($\overline{n}$) for the number of potential members and that this constraint might bind. Furthermore, when the parties are heterogeneous, or with minimum participation constraints, a number ($\underline{n}$) of them may participate regardless of the procedure. When these constraints are accounted for, I show that P&R is preferred if and only if $\overline{n}$ is large and/or $\underline{n}$ is small.

The optimal contract duration in this model results from a novel trade-off: a long-term contract is unattractive because, after the parties have invested in the capacity to produce renewable energy, it becomes optimal to negotiate still more ambitious pledges. A short-term contract, however, creates a hold-up problem when the parties anticipate how their investments will influence the

next bargaining outcome. The optimal term trades off these two concerns, but this trade-off is the same under P&R as under NBS, and I find that it is independent of the number of participants in this model.

While most of the paper considers Markov-perfect equilibria (MPEs) and abstracts from non-compliance, I end the analysis by considering when a party would be tempted to defect after the agreement has been made. I show that P&R is more likely than the NBS to be self-enforcing. Equivalently, if the agreement must be self-enforcing for exogenous reasons, the contributions, or the weights on others' payoffs, cannot be too large.

I do not force the model to be more complicated than is necessary. However, some readers may react to assumptions regarding the timing of the game, the choice of policy instruments and the tools available at the bargaining stage. To deal with these concerns, the robustness section shows that the model can be generalised in ten such directions and I explain why the results survive in all of them.

Empirically, the predictions are consistent with several of the differences between the Kyoto Protocol and the Paris Agreement. Bargaining theory predicts that, under Paris, the weights on others' payoffs are smaller both because of the P&R procedure and because the emission cuts are not legally binding (in contrast to Kyoto). With this, it is consistent with the model that few countries made commitments under Kyoto, while participation was much larger in Paris. The model is also consistent with the development from Kyoto to Paris: in the 1990s, there were a large number of developing countries that could not be expected to contribute much to a global climate policy. Over the last 20 years, some of these have become emerging economies that potentially have important roles to play. The number of relevant potential parties, $\overline{n}$, has therefore increased. During the same period, seven countries that initially signed the Kyoto Protocol declared that they did not intend to contribute to Kyoto's second commitment period (IPCC, 2014, p. 1025). This can be interpreted as a smaller $\underline{n}$. Either (or both) of these developments should make P&R relatively more attractive for every participant, according to the theory. As argued in Section 6, the theory is even consistent with the disagreement between the North and the South when it came to the choice of procedure, and that the length of the commitment period was the same (five years) for both treaties.

After explaining my contribution to the literature, Section 2 presents the model. Section 3 derives equilibrium investment levels, emission cuts and participation levels before discussing the choice of bargaining procedure and the optimal commitment period length. In Section 4, I discuss compliance and provide two micro-foundations for the assumption that the weights on others' payoffs are smaller under Paris. Ten generalisations are analysed in Section 5, before Section 6 relates the theoretical predictions to the empirical facts. After a brief concluding section, the Appendix presents all proofs.

1. The Literature

The dynamic climate change game below draws on standard assumptions introduced by Dutta and Radner (2004; 2006; 2020), Harstad (2012; 2016) and Battaglini and Harstad (2016), although the chosen functional forms here are different, making the analysis below especially tractable. More importantly, *none* of the above papers compare or derive the consequences of alternative bargaining outcomes.

Dutta and Radner (2020) and Caparrós (2020) justified an aspect of the Paris Agreement by showing how its Green Climate Fund can lead to efficiency. Thus, they complemented the present

paper, which instead focuses on the game between developed countries, with neither transfers nor conditionality.

The coalition formation game below is the standard one when modelling collusion (d'Aspremont *et al.*, 1983; Bloch, 2018) and environmental coalitions (Hoel, 1992; Carraro and Siniscalco, 1993; Barrett, 1994). The typical prediction is that the coalition size is very small. This prediction is inconsistent with the fact that real-world coalitions can be quite large. This inconsistency is referred to as a 'paradox' by Kolstad and Toman (2005) and Nordhaus (2015).

Dynamic models can explain large coalitions (see the surveys by Calvo and Rubio, 2012; Caparrós, 2016) depending on the contractual environment (Battaglini and Harstad, 2016), the lobby (Marchiori *et al.*, 2017), re-election concerns (Battaglini and Harstad, 2020), beliefs (Karp and Sakamoto, 2021) or the willingness to avoid delays (Kovác and Schmidt, 2021). These papers do not explain why many more countries contribute to the Paris Agreement than to the Kyoto Protocol, however.

The narrow-but-deep versus broad-but-shallow trade-off is well known; see Schmalensee (1998), Barrett (2002) or Aldy *et al.* (2003). Finus and Maus (2008) showed that modest contributions can increase participation and payoffs. They assumed the bargaining outcome and a static game. These assumptions have been criticised by political scientists: arguably, the trade-off vanishes with heterogeneous policies (Gilligan, 2004) or endogenous enforcement (Bernauer *et al.*, 2013). My analysis rediscovers the trade-off when the bargaining game and compliance are micro-founded. In contrast to Schmalensee (1998), who recommended 'broad, then deep', the results in this paper are consistent with the reverse, factual development from the deep-but-narrow Kyoto Protocol to the broad-but-shallow Paris Agreement.

Finally, I contribute to the literature on self-enforcing agreements; see, for example, Barrett (1994; 2002); Dutta and Radner (2004; 2006); Harstad *et al.* (2019; 2022) and the references therein. I show when and why certain procedures, such as pledge and review, are more likely than others to be self-enforcing.

2. Model and Benchmark Results

2.1. *Dynamic Game with Contributions and Investments*

The model describes a situation in which a set of parties can contribute to a public good as well as invest in their future capacities to contribute. In equilibrium, the negotiated contribution levels will influence how much the parties will invest, but past investments will also influence the future contribution levels. Although the model can be applied to other public good settings, it fits especially well to analyse climate policies. As required by the Paris Agreement (Art. 4.9), 'Each Party shall communicate a nationally determined contribution every five years.' Apparently, 'The idea is that this short time frame would give countries the opportunity to regularly capture scientific and technological developments in their official targets.'[2] The Stern Review (Stern, 2006) also pointed out that new technology would be crucial to mitigate climate change. However, the treaties establish that 'technology needs must be nationally determined, based on national circumstance and priorities' (Section 114 of the 2010 Cancun Agreement). For the model to be consistent with this practice, emission cuts are assumed to be

[2] https://www.carbonbrief.org/explainer-the-ratchet-mechanism-within-the-paris-climate-deal.

negotiable and contractible, while technology investments are not. (The assumption is relaxed in Section 5.)

In each period t, the utility for a party is the sum of three parts. First, if each party i contributes or abates the quantity $q_{i,t}$, the sum of abatements has the value $a\sum_{i\in N} q_{i,t}$ to each party. This linearity assumption is made for simplicity, but it is common also because it is a reasonable approximation when it comes to climate change. As Golosov *et al.* (2014, p. 78) wrote, for example: 'Linearity is arguably not too extreme a simplification, since the composition of a concave S-to-temperature mapping with a convex temperature-to-damage function may be close to linear.'

An additional benefit of this linearity is that we can easily allow for a stock of greenhouse gases that accumulates over time, without changing the analysis, because a can be interpreted as the present discounted cost of emitting another unit of emission into the atmosphere, when we anticipate that this unit may contribute to climate change for decades. To see this, suppose that party i emits $g_{i,t}$ and that the pollution stock is $G_t = \sigma G_{t-1} + \sum_{i\in N} g_{i,t}$, where $\sigma \in [0, 1]$ measures the fraction of the past stock that survives to the next period. If parameter $h > 0$ measures each party's per-period marginal environmental harm from stock G_t then the present discounted harm of another unit of emission is $h/(1-\sigma\delta)$ for each party. Consequently, $a \equiv h/(1-\sigma\delta)$ measures the present discounted benefit from abating one unit.

The second term in the utility function specifies the cost of contributing to the public good. For example, suppose that a country can consume energy from both fossil fuels ($g_{i,t}$) and renewables ($R_{i,t}$). If the total consumption of energy is less than i's bliss point, $\overline{g}_{i,t}$, then i may experience a disutility that is quadratic in the difference: $b(\overline{g}_{i,t} - [g_{i,t} + R_{i,t}])^2/2$. This disutility can be written as $b(q_{i,t} - R_{i,t})^2/2$, when $q_{i,t}$ represents a cut in emissions relative to i's bliss point (i.e., when $q_{i,t} \equiv \overline{g}_{i,t} - g_{i,t}$).[3]

Of course, for other public good situations also, it can be especially costly for i to contribute a lot relative to i's capacity level, as represented by stock $R_{i,t}$.

Each party can over time add to the capacity $R_{i,t}$ by investing $r_{i,t}$. The investment cost is assumed to be convex and quadratic. The marginal investment cost is $\underline{c}_{i,t} + cr_{i,t}$, where $\underline{c}_{i,t}$ can be heterogeneous, time dependent and negative as well as positive. The investment cost constitutes the third term in the per-period utility function,

$$u_{i,t} = a\sum_{j\in N} q_{j,t} - \frac{b}{2}(q_{i,t} - R_{i,t})^2 - \frac{c}{2}\left(\frac{\underline{c}_{i,t}}{c} + r_{i,t}\right)^2, \quad \text{where} \quad R_{i,t+1} = R_{i,t} + r_{i,t}, \qquad (1)$$

and where a, b and c are positive constants. The parties can have heterogeneous $\overline{g}_{i,t}$, $\underline{c}_{i,t}$ and initial technology levels ($R_{i,1}$). For simplicity, and to discipline the model, the other parameters are assumed to be constant over time and across the parties. Because I abstract from technological spillovers, the $r_{i,t}$ may be best interpreted as renewable energy capital or infrastructure, rather than investments in knowledge.

[3] For this particular interpretation of the model, $g_{i,t} = B_{i,t} - q_{i,t}$ could be negative if $q_{i,t}$ is very high compared to $B_{i,t}$. I do not impose any constraint $g_{i,t} \geq 0$ because (a) for simplicity, (b) $g_{i,t} \geq 0$ will not bind if $B_{i,t}$ is growing sufficiently fast over time, (c) $g_{i,t} < 0$ is in reality feasible with carbon-capture and storage technologies and (d) it should be possible to interpret $q_{i,t}$ as (unbounded) contributions to a public good, more generally. See Harstad (2012) for how one may deal with the constraint $g_{i,t} \geq 0$ in a similar (although somewhat different) model without affecting the results qualitatively.

When δ is the common discount factor, party i intends to maximise the following continuation value at each time t:

$$V_{i,t} = u_{i,t} + \delta V_{i,t+1} = \sum_{\tau=t}^{\infty} \delta^{\tau-t} u_{i,\tau}.$$

2.2. *Business as Usual and First Best*

As there is a large number of subgame-perfect equilibria (SPEs) in dynamic games, it is common to restrict attention to Markov-perfect strategies when there are stocks in the game. This refinement pins down the unique equilibrium that is the limit of the unique SPE if the time horizon were finite but approached infinity.

In the MPE without any treaty, i.e., the 'business as usual' (BAU) equilibrium, all the $r_{i,t}$ and $q_{i,t}$ are decided on simultaneously and non-cooperatively. Because of the constant marginal cost of emission, the pollution stock is not payoff relevant and the MPE is unique.[4] In every period t, when i takes as given $R_{i,t}$, the marginal abatement cost equals the marginal benefit for party i:

$$b(q_{i,t}^{BAU} - R_{i,t}) = a \quad \Longleftrightarrow \quad q_{i,t}^{BAU} = R_{i,t} + \frac{a}{b}.$$

Consequently, $r_{i,t}$ does not influence i's actual contribution cost, but only i's contribution level in every future period. Party i's preferred investment level is thus

$$r_{i,t}^{BAU} = \frac{\delta}{1-\delta}\frac{a}{c} - \underline{c}_{i,t},$$

if we assume that $\underline{c}_{i,t} < [\delta/(1-\delta)](a/c)$. With this, it is straightforward to derive party i's continuation value in BAU, $V_{i,t}^{BAU}$.[5]

The first-best outcome is given by the exact same equations if a is just replaced by na. In both cases, the second-order conditions trivially hold.

2.3. *Pledges*

Now, consider the situation that arises after the parties have agreed to contribute *more* than the BAU levels. In particular, suppose that i has agreed to contribute $x_i \geq 0$ units, beyond i's BAU level, for each of the next T periods. Clearly, the commitment x_i is payoff relevant and it might motivate i to invest $y_{i,t}$ units in addition to the BAU level. Total contributions and investments can be written as follows:

$$q_{i,t} \equiv q_{i,t}^{BAU} + x_i \quad \text{and} \quad r_{i,t} \equiv r_{i,t}^{BAU} + y_{i,t}. \tag{2}$$

[4] I start by deriving party i's unique best response, assuming that other parties do not condition *their* strategies on any stock, and find that i's strategies do not depend on the stocks either. Consequently, Markov-perfect strategies are not conditioned on any stock. As Maskin and Tirole (2001, p. 202) wrote: 'Markov strategies are the simplest strategies (i.e., the strategies measurable with respect to the coarsest partition and hence dependent on the fewest variables) that are consistent with rationality in the sense that, if the other players make their strategies measurable with respect to some [even] coarser partition [of the history] it would *not* always be optimal for a player to make his or her choice between any two given continuation strategies measurable with respect to [that partition].'

[5] As proven in the Appendix,

$$V_{i,t}^{BAU} = \frac{a}{1-\delta}\sum_{j\in N}\left(R_{j,t} - \sum_{\tau=t}^{\infty}\delta^{\tau+1-t}\underline{c}_{i,\tau}\right) + \frac{a^2}{1-\delta}\left(\frac{1}{b} + \frac{1}{c}\left[\frac{\delta}{1-\delta}\right]^2\right)\left(n - \frac{1}{2}\right).$$

It will turn out to be most convenient to focus on the choices of x_i and $y_{i,t}$ (of course, these decisions pin down $q_{i,t}$ and $r_{i,t}$, given BAU).

As soon as the parties have agreed on $\mathbf{x} = (x_1, \ldots, x_n)$, i faces a continuation value that can be written as $V_{i,t}^{BAU} + U_i(\mathbf{x})$. The function $U_i(\mathbf{x})$ is i's bargaining surplus, relative to BAU, which will be characterised in Section 3.

2.4. *Nash versus Pledge-and-Review Bargaining*

If bargaining were frictionless and side transfers available, the Coase theorem would suggest that all the x_i would be efficient, regardless of the bargaining procedure. Explicit side transfers are rare in international politics, however. Thus, the procedure plays a role.

A standard bargaining solution in applied theory is the NBS, predicting that every x_i^* is the outcome of the following maximisation problem:

$$x_i^* = \arg\max_{x_i} \prod_{j \in N} U_j(x_i, \mathbf{x}_{-i}^*).$$

A generalisation of the NBS is to let x_i^* maximise a generalised or *asymmetric* Nash product, where the weight on others' payoffs can be smaller:

$$x_i^* = \arg\max_{x_i} \prod_{j \in N} U_j(x_i, \mathbf{x}_{-i}^*)^{w_i^j}, \quad \text{where} \quad w_i^j / w_i^i = w \in [0, 1) \quad \text{for } j \neq i. \tag{3}$$

When all utility levels are the same in equilibrium (i.e., when the situation is symmetric), the first-order condition (f.o.c.) for (3) is equivalent to the f.o.c. for

$$x_i^* = \arg\max_{x_i} \left[U_i(x_i, \mathbf{x}_{-i}^*) + w \sum_{j \in N \setminus i} U_j(x_i, \mathbf{x}_{-i}^*) \right]. \tag{4}$$

While the NBS has often been used to characterise bargaining outcomes, such as the commitments under the Kyoto Protocol, the discussion provided in the introduction suggests that, for P&R, $w < 1$. The formal analysis in Subsection 4.1, in part drawing on my companion paper (Harstad, 2023), verifies that there are two reasons for why the weight on others' payoffs is smaller (say, $\underline{w}$) under P&R than it was for the Kyoto Protocol (say, $\overline{w}$). The subsequent analysis will investigate and compare the consequences of $\underline{w} \in (0, 1)$ versus $\overline{w} \in (\underline{w}, 1]$.

2.5. *Participation*

There are alternative ways of modelling participation, but the standard approach is simple and by adopting it, I can clarify my contribution to the literature. According to Nordhaus (2015, p. 1344), 'the standard approach in environmental economics' when modelling coalitions begins with a participation stage at which every country, $i \in \{1, \ldots, \overline{n}\}$, decides whether to participate in the coalition. These decisions are made simultaneously and everyone expects that participants will negotiate pledges according to (4). Given the restriction to MPEs, free riders will simply follow their dominant BAU strategy and set $x_i = 0$.

Countries are more heterogeneous in reality than permitted in the model above. A simple way of capturing heterogeneity is to let $\underline{n}$ parties be committed in that they participate regardless of what other countries do. The reason these parties are committed can be outside the model, but one may think of existing treaties on non-climate issues such as trade or regulatory politics. To be

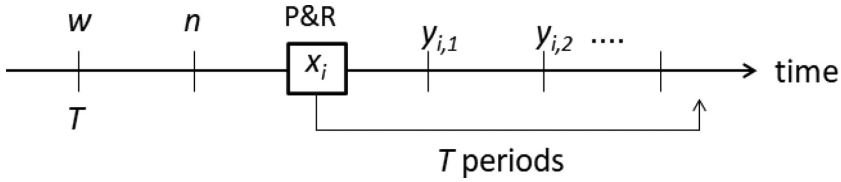

Fig. 1. *The Timing.*

specific, European Union member countries cannot easily opt out of an environmental agreement unilaterally.

There can also be a minimum participation level for other reasons. Most international treaties specify minimum participation thresholds that must be met for the treaty to enter into force. This threshold was the same for the Kyoto Protocol and the Paris Agreement. In this model, the effect of such a threshold, if we refer to it as $\underline{n}$, will be similar to the effect of committed countries. After all, when the threshold binds, none of the $\underline{n}$ participants prefer to free ride given that the consequence will be the BAU outcome.

2.6. *Timing*

Figure 1 illustrates the timing. After the bargaining game (characterising w) is determined, countries choose between participation and free riding, determining n. Then, emission cuts are negotiated. In each of the T periods after an agreement has been made, the pledge x_i pins down $q_{i,t}$ but party i is free to set $r_{i,t}$ or, equivalently, $y_{i,t}$.

In Section 3, I first describe equilibrium investments, given the pledges, and equilibrium pledges and payoffs, given the coalition size. The optimal T is discussed in Subsection 3.5. (It turns out that it does not matter whether T is set before or after w.) Section 4 provides micro-foundations for w and considers compliance constraints. In Section 5, I discuss extensions, including alternative timings of the game.

The below results hold whether or not the parties negotiate new pledges after the present T-period commitment period. To distinguish the two cases, the index $\iota \in \{0, 1\}$ takes the value of 1 if a new commitment period will be negotiated every T period, but $\iota = 0$ if one returns to BAU after the current T-period commitment period. (Note that we have $\iota = 1$ for the Paris Agreement because new pledges are supposed to be decided on every five years. This fact will be rationalised in Subsection 4.2; see footnote 13.)

3. Equilibrium

3.1. *Equilibrium Investments*

After the pledges have been agreed on, party i's problem is to choose the investment levels over the next T periods. This boils down to a standard optimal control problem, solved in the Appendix. The exact solution for the investment levels is presented here.

PROPOSITION 1. *For each $i \in N$, $t \in \{1, .., T\}$ and $\iota \in \{0, 1\}$, equilibrium investments increase in x_i:*

$$y_{i,t} = x_i(l_1 m_1^{t-1}[1 - m_1] - l_2 m_2^{t-1}[m_2 - 1]),$$

where

$$m_1 \equiv \frac{1}{2}\left(\frac{1}{\delta}+1+\frac{b}{c}\right)-\frac{1}{2}\sqrt{\left(\frac{1}{\delta}+1+\frac{b}{c}\right)^2-\frac{4}{\delta}} \in (0,1),$$

$$m_2 \equiv \frac{1}{2}\left(\frac{1}{\delta}+1+\frac{b}{c}\right)+\frac{1}{2}\sqrt{\left(\frac{1}{\delta}+1+\frac{b}{c}\right)^2-\frac{4}{\delta}} > 1,$$

$$l_1 \equiv \frac{m_2^{T-1}(m_2-1)}{m_1^{T-1}(1-m_1)+m_2^{T-1}(m_2-1)} \in (0,1),$$

$$l_2 \equiv \frac{m_1^{T-1}(1-m_1)}{m_1^{T-1}(1-m_1)+m_2^{T-1}(m_2-1)} = 1-l_1 \in (0,1).$$

Naturally, if i has committed to contribute a lot, in that x_i is large, then i invests more. It is easy to check that $y_{i,t}$ increases in T and decreases in t. In the final period, a party invests $y_{i,T}=0$, exactly the same amount as in BAU. (This holds whether one expects to negotiate new pledges in the next period ($\iota=1$) or not ($\iota=0$).) The intuition is related to the hold-up problem: one more technology unit in the next period can—without any other change in investment or contribution cost—raise the total contribution level by one unit then and forever after. The party that invested captures $1/n$ of this benefit, just as in BAU (see Harstad, 2016 and Battaglini and Harstad, 2016). This intuition also explains why the equilibrium investment at any point in time is the same whether future agreements are expected (i.e., $\iota=1$) or not ($\iota=0$). An important implication is that in every period in which the parties have not yet agreed to any pledge, contribution and investment levels are just as in BAU: $x_i=0$ and $y_{i,t}=0$.

3.2. *Equilibrium Pledges*

Conveniently, Proposition 1 states that the level of investments and thus technologies will be linear functions of x_i. We can substitute these functions into i's utility function and write party i's continuation value (i.e., the present discounted value of the future utility levels) as a linear-quadratic function, $V_{i,1}(\mathbf{x})$, where $\mathbf{x}\equiv(x_1,\ldots,x_n)$. Given the benchmark continuation value without an agreement, $V_{i,1}^{BAU}$, we are especially interested in the additional payoff with the pledges: $U_i(\mathbf{x})\equiv V_{i,1}(\mathbf{x})-V_{i,1}^{BAU}$. The additional payoff $U_i(\mathbf{x})$ turns out to be a simple linear-quadratic function, although with parameters α, β and γ being complicated functions of a, b, c, δ and T, as proven in the Appendix.[6]

LEMMA 1. *Party i's continuation value, relative to BAU, can be written as*

$$U_i(\mathbf{x})=\alpha\sum_{j\neq i}x_j-\frac{\beta}{2}x_i^2+\gamma, \tag{5}$$

[6] The additional payoff $U_i(\mathbf{x})$ ends up being symmetric for all parties because the heterogeneous $\overline{g}_{i,t}$, $\underline{c}_{i,t}$ and $R_{i,1}$ cancel when $V_{i,1}^{BAU}$ is subtracted from i's continuation value. In contrast, if a, b or c were heterogeneous, $U_i(\mathbf{x})$ would not be symmetric, and thus (3) would not simplify to (4). If technological spillovers were permitted, $\partial U_i(\mathbf{x})/\partial x_j$ would not be constant.

where

$$\alpha \equiv \frac{a}{1-\delta}[1-\delta^T(l_1 m_1^{T-1} + l_2 m_2^{T-1})],$$

$$\beta \equiv \sum_{t=1}^{T} \delta^{t-1}[b(l_1 m_1^{t-1} + l_2 m_2^{t-1})^2 + c(l_1 m_1^{t-1}[1-m_1] - l_2 m_2^{t-1}[m_2 - 1])^2],$$

$$\gamma \equiv \delta^T U_i(\mathbf{x}^*)\iota.$$

When we combine (4) with Lemma 1, we obtain the following result.

LEMMA 2. *For P&R ($w = \underline{w}$) and the NBS ($w = \overline{w}$), the equilibrium pledge is*

$$x_i^* = w\,(n-1)\,\alpha/\beta. \tag{6}$$

The smaller w is, the smaller are the x_i^* and, according to Proposition 1, the smaller are the investment levels. These effects make parties worse off than in the situation in which $w = 1$. By combining (5) and (6), we see that $U_i(\mathbf{x}^*)$ increases in w when $w \in (0, 1)$.

PROPOSITION 2. *For any given n, with a smaller w (as with $\underline{w} < \overline{w}$), contributions are lower, investments are lower and so is welfare:*

$$U_i(\mathbf{x}^*) = \frac{\alpha^2(n-1)^2}{\beta(1-\delta^T\iota)} w\left(1 - \frac{w}{2}\right). \tag{7}$$

3.3. *Participation*

It is most natural (and common) to focus on pure-strategy equilibria at the participation stage, and doing so pins down the equilibrium coalition size, n. I start by ignoring the constraint $n \leq \overline{n}$ as well as a possible minimum participation threshold, $\underline{n}$, but these constraints are extensively discussed in the next subsection. I also begin by assuming that the participation decision is made once and for all. Section 5 explains why the results continue to hold when this assumption is relaxed.

Because coalition members end up contributing more than the level that would maximise their own utility, there is a cost of participating in the coalition. For a member to be willing to participate, the benefit of participating must outweigh this cost. The benefit of participating is that other participants will internalise (a fraction w of) the utility of one additional coalition member.

For each of the n participants, the equilibrium payoff (in addition to the BAU payoff) is given by (7). If one of these parties instead free rides, the free rider's additional payoff will be $\alpha(n-1)w(n-2)\alpha/\beta(1-\delta^T\iota)$, because each of the other $n-1$ parties will now commit to contribute $w(n-2)\alpha/\beta$, again and again, every T period (if $\iota = 1$). By comparison, participation is beneficial if

$$U_i(\mathbf{x}^*) = \frac{\alpha^2(n-1)^2}{\beta(1-\delta^T\iota)} w\left(1 - \frac{w}{2}\right) \geq \frac{\alpha^2(n-1)(n-2)}{\beta(1-\delta^T\iota)} w \quad \Longleftrightarrow \quad n \leq 1 + \frac{2}{w}. \tag{8}$$

The size n cannot be too great because then individual contributions would be so large and so costly that free riding would be preferable. For a coalition to be stable, (8) must hold for the equilibrium n. At the same time, we must have $n' > 1 + 2/w$ for every $n' > n$, as otherwise,

non-members would also like to participate. To characterise the equilibrium n, it is useful to employ the function $\lfloor \cdot \rfloor$, mapping its argument to the largest weakly smaller integer.

PROPOSITION 3. *With a smaller w (as with $\underline{w} < \overline{w}$), the coalition size is larger:*

$$n = \lfloor 1 + 2/w \rfloor .$$

Note that $n = 3$ if $w = 1$, as when applying the NBS. This 'small-coalition paradox' is well known in the literature, which also discusses the trade-off between 'narrow-but-deep' versus 'broad-but-shallow' coalitions (see the literature review). If, with P&R, w is small, a coalition member is not expected to contribute a lot. Lower contributions reduce both the cost and the benefit of participating. The first effect dominates because the cost is a strictly convex function of the contribution, while the benefit function is concave (linear). Thus, when w is small, participation is attractive for a larger set of n.

As the number of participants must be an integer, n is a step function of w. Approximately,

$$n \approx n(w) \equiv 1 + 2/w. \tag{9}$$

In fact, one may argue that (9) must hold with equality at the optimal bargaining game. To see this, let W be defined as the set of w, $w \in [0, 1]$, such that $1 + 2/w$ is an integer. Thus, $W = \{\ldots, w_5, w_4, w_3\}$, where $w_n = 2/(n-1)$, $n \in \mathbb{N}$ and $n \geq 3$. If $w \in [0, 1]$, but $w \notin W$, then $w \in (w_{n+1}, w_n)$ for some $n \geq 3$. In line with Proposition 2, contributions, investments and all payoffs are larger if w is increased to w_n. In other words, $w \in (w_{n+1}, w_n)$ is (Pareto) dominated by w_n. Hence, an optimal w must satisfy $w \in W$.

With (9), the product $(n-1)w$ is a constant that is pinned down because a member must be indifferent between free riding and participating.

3.3.1. *Equilibrium pledges—revisited*

When $(n-1)w$ stays constant as w is reduced, $x_i^* = (n-1)w\alpha/\beta = 2\alpha/\beta$ also remains constant, and so does every investment level $y_{i,t}$.[7] When the individual contributions are invariant in w, while n is decreasing in w, the sum of payoffs will be larger when w is small. A participant's payoff is also larger when w is small: this is evident when the endogenous n, as described by (9), is combined with the utility (7). This gives[8]

$$U_i^* = \frac{4\alpha^2}{\beta(1-\delta^T\iota)}\left(\frac{1}{w} - \frac{1}{2}\right), \qquad w \in W. \tag{10}$$

[7] The result that x_i stays unchanged when n varies endogenously with w hinges on the functional forms. If, for example, the continuation value had ended up being

$$\alpha \sum_{j \neq i} x_j - \frac{\beta}{2} x_i^{\varphi} \quad \text{with } \varphi > 1,$$

then one can show that: n always decreases in w; x_i decreases in w if $\varphi < 2$ but increases in w if $\varphi > 2$; U_i^* always decreases in w when $w \in [0, 1]$.

[8] If $w \notin W$, so that $w \in (w_{n+1}, w_n)$ for some n, the payoff is given by (10) minus the loss associated with a smaller $w < w_n$, given n. This loss follows from (7):

$$\frac{\alpha^2(n-1)^2}{\beta(1-\delta^T\iota)}\left[\left(w_n - \frac{w_n^2}{2}\right) - \left(w - \frac{w^2}{2}\right)\right].$$

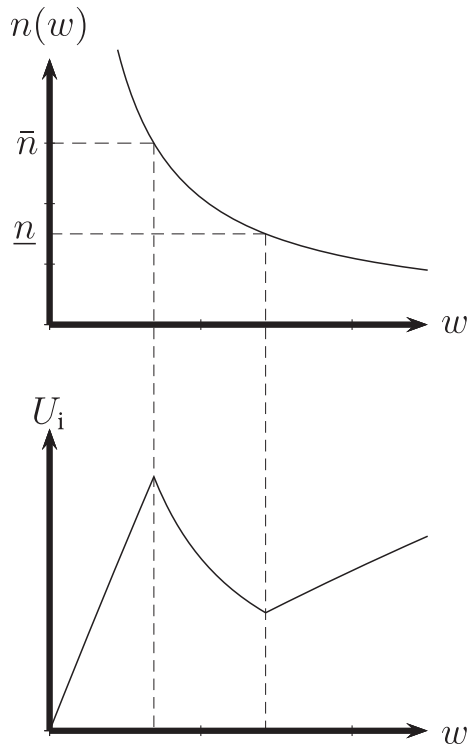

Fig. 2. *Participation and Participants' Payoffs are Strictly Decreasing in w only when* $n(w) \in (\underline{n}, \overline{n})$.

COROLLARY 1. *When participation is endogenous, and approximated by n(w), Proposition 2 is reversed: with a smaller* $w \in W$ *(e.g., with* $\underline{w} < \overline{w}$*), aggregate contributions are larger, aggregate investments are larger and so is welfare.*

Finus and Maus (2008) also found that 'modesty' can increase participation, and thus contributions and payoffs. However, they abstracted from investments, dynamics and micro-foundations for the modesty.

3.4. *When to Choose Pledge and Review*

Figure 2 illustrates that, with the constraints $n \leq \overline{n}$ and $n \geq \underline{n}$, payoffs are non-monotonic in w. This section discusses each constraint separately, before combining them.

The effect of maximum participation. There is a limit, $\overline{n}$, for how large the coalition can be. One may interpret $\overline{n}$ as the number of countries in the world or, alternatively, as the number of countries that are of significance.

If $\overline{n} < n(\overline{w}) < n(\underline{w})$, where $n(\cdot)$ is defined by (9), then both bargaining games (characterised by $\underline{w}$ or $\overline{w}$) induce full participation. In this case, $\overline{w}$ is preferable, according to Proposition 2. If, instead, $n(\overline{w}) < n(\underline{w}) < \overline{n}$, the upper boundary on n is non-binding and $\underline{w}$ is preferable, according to the corollary. A trade-off arises when $n(\overline{w}) < \overline{n} < n(\underline{w})$, because then participation is larger, but individual contributions smaller, when w is small. In this case, a sufficiently large $\overline{n}$ is necessary to ensure that a participant's payoff is larger under $\underline{w}$.

The exact condition follows when comparing a participant's utility, as given by (7), for the two cases. The payoff is larger when $w = \underline{w}$ than when $w = \overline{w}$ if

$$\frac{\alpha^2(\overline{n}-1)^2}{\beta(1-\delta^T\iota)}\underline{w}^2\left(\frac{1}{\underline{w}}-\frac{1}{2}\right) > \frac{\alpha^2(n(\overline{w})-1)^2}{\beta(1-\delta^T\iota)}\overline{w}^2\left(\frac{1}{\overline{w}}-\frac{1}{2}\right) \quad \Longrightarrow \quad \frac{\overline{n}-1}{n(\overline{w})-1} > \Omega,$$

$$\text{where} \quad \Omega \equiv \sqrt{\frac{\overline{w}(1-\overline{w}/2)}{\underline{w}(1-\underline{w}/2)}} \in \left(1, \frac{\overline{w}}{\underline{w}}\right).$$

The effect of minimum participation. The minimum participation threshold $\underline{n}$ is relevant only if $\underline{n} > n(\overline{w})$. If $\underline{n} > n(\underline{w})$ also, the number of participants is always $\underline{n}$ and then the larger $\overline{w}$ is better than $\underline{w}$, according to Proposition 2.

To isolate the trade-off associated with $\underline{n}$, let $n(\overline{w}) < \underline{n} < n(\underline{w}) < \overline{n}$. In this case, $\underline{n}$ parties participate under $\overline{w}$, while participation under $\underline{w}$ is given by $n(\underline{w})$. By comparison, a participant's payoff can be larger under $\underline{w}$ if and only if $\underline{n}$ is sufficiently small. The exact condition follows when we use the utility function (7) to compare the two cases. The payoff is larger when $w = \underline{w}$ than when $w = \overline{w}$ if

$$\frac{\alpha^2(n(\underline{w})-1)^2}{\beta(1-\delta^T\iota)}\underline{w}^2\left(\frac{1}{\underline{w}}-\frac{1}{2}\right) > \frac{\alpha^2(\underline{n}-1)^2}{\beta(1-\delta^T\iota)}\overline{w}^2\left(\frac{1}{\overline{w}}-\frac{1}{2}\right) \quad \Longrightarrow \quad \frac{n(\underline{w})-1}{\underline{n}-1} > \Omega.$$

The preferred bargaining game. Clearly, it is possible that both the minimum and the maximum participation levels bind at the same time. This happens if $n(\overline{w}) < \underline{n} < \overline{n} < n(\underline{w})$. In this case, there is full participation under $\underline{w}$, but only $\underline{n}$ parties participate under $\overline{w}$. In this situation, $\underline{w}$ is preferred when $\overline{n}$ is large and $\underline{n}$ is small. From (7), we get that the exact condition is

$$\frac{\alpha^2(\overline{n}-1)^2}{\beta(1-\delta^T\iota)}\underline{w}^2\left(\frac{1}{\underline{w}}-\frac{1}{2}\right) > \frac{\alpha^2(\underline{n}-1)^2}{\beta(1-\delta^T\iota)}\overline{w}^2\left(\frac{1}{\overline{w}}-\frac{1}{2}\right) \quad \Longrightarrow \quad \frac{\overline{n}-1}{\underline{n}-1} > \Omega.$$

The three conditions can be combined in the following way.

PROPOSITION 4. *Everyone who participates when $w = \overline{w}$ prefers to switch to P&R and $w = \underline{w} < \overline{w}$ if $\overline{n}$ is large while $\underline{n}$ is small. The exact condition is*

$$\frac{\min\{\overline{n}-1, 2/\underline{w}\}}{\max\{\underline{n}-1, 2/\overline{w}\}} > \Omega. \tag{11}$$

This condition is drawn as the solid line in Figure 3. If there is a larger number of potential parties, or if fewer of them are committed to participate, we move in the direction of the arrow in the figure. Then, the 'shallow' agreement ($\underline{w}$) becomes preferred by all participants though the 'deep' agreement was preferred given a smaller number of potential parties or a larger number of committed parties.

Interestingly, there may be a disagreement between the insiders and the outsiders regarding what procedure to choose. By using the same methodology as above, one can show that the *uncommitted* countries prefer P&R ($\underline{w}$) only when

$$\min\{\overline{n}-1, 2/\underline{w}\} > \sqrt{\frac{\max\{\overline{w}(\underline{n}-1)\underline{n}, 4/\overline{w}+2\}}{\underline{w}(1-\underline{w}/2)}}, \tag{12}$$

which is stronger than (11), although the comparative statics are similar. In other words, the original set of (committed) participants prefer to switch to P&R too soon, that is, for a larger

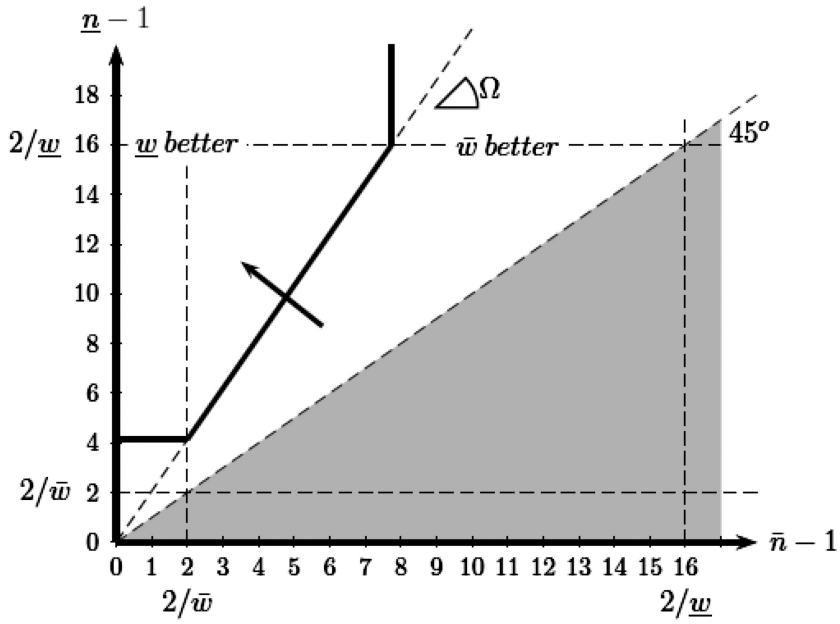

Fig. 3. *Participants Prefer to Switch to Pledge and Review ($\underline{w}$) above the Solid Line.*

set of parameters than the set under which such a switch increases global welfare. Analogously, if the new potential members were pivotal in the decision on treaty design, they would accept P&R too late or too seldom, relative to the decision that is optimal when the original members' payoffs are taken into account.

3.5. *The Commitment Period Length*

The results above hold for any commitment period length. The optimal T, from the participants' point of view, trades off two effects. On the one hand, the shorter the length of the commitment period, the lower the equilibrium investments at every point in time. This comparative static can be seen from Proposition 1 and it was explained above by the classic hold-up problem. On the other hand, with a large T, the x_i will soon be outdated because it will be optimal to deepen the cuts and the contributions after investments have accumulated.[9] The combined trade-off is new to the literature.[10]

The trade-off when it comes to deciding on T is independent of w and n in the model above. When n is exogenous, a party's payoff is given by (7). When n is instead endogenous, a party's

[9] The pledges do permit decreasing pollution levels, as the x_i are defined as cuts relative to the BAU outcome. However, countries invest more, given the pledges, than they do in the BAU outcome, and the importance of these additional investments accumulate over time. Section 5 explains that the results continue to hold if the pledges are functions of time or of investments.

[10] On the one hand, Harris and Holmstrom (1987) observed that a small T is beneficial because it permits a rigid contract to be updated when the external environment changes. On the other hand, the hold-up problem associated with a small T is recognised by, for example, Beccherle and Tirole (2011) and Harstad (2016). These two effects have never been combined, as far as I know.

payoff is given by (10). Every participant's preferred T is the same in either case:

$$T^* = \arg\max_T U_i(\mathbf{x}^*) = \arg\max_T U_i^* = \arg\max_T \frac{\alpha^2}{\beta(1-\delta^T\iota)}$$

with α and β functions of T, as described by Lemma 1. This expression is complicated and depends on many of the model's parameters. For example, the optimal T is larger if $\iota = 0$ than if $\iota = 1$ because, in the first case, when no new commitment period will replace the current one, then the duration of cooperation is essentially given by T. However, the optimal T does not depend on parameter a. The intuition is that a larger a increases *all* benefits and costs of abatements, without altering the trade-off, described above. Thus, even if a changes over time, T^* does not change, in this model.

More interestingly, T^* does not depend on w and n.

PROPOSITION 5. *The optimal commitment period length T^* is invariant in w and the same with $\underline{w}$ and $\overline{w}$, regardless of whether n and w are endogenous or exogenous.*

4. Pledge and Review: Foundations and Enforcement

Compared to Kyoto, the Paris Agreement is distinct in multiple respects. Tubiana and Guérin (2020, p. 112) wrote that 'the agreement is largely bottom-up in nature, because it is based on self-determined contributions' and it lacks legally binding commitments.[11] In this section, I analyse the consequences of emission cuts that are (1) unconditional versus conditional on other countries' emission cuts and (2) legally binding versus not legally binding. According to my formal analysis, both differences suggest that the weights on other parties' payoffs are smaller under Paris-style P&R ($\underline{w}$) than under the procedure used for the Kyoto Protocol ($\overline{w}$). Furthermore, the two differences are interrelated, and I find they go naturally together.

4.1. *A Theory of Pledge-and-Review Bargaining*

The simplicity of the NBS is one reason for why it is popular and reasonable in applications. In addition, the NBS relies on an appealing set of axioms (Nash, 1950) and it is the outcome of non-cooperative bargaining games, such as the Nash demand game (Nash, 1953). The Rubinstein (1982) alternating-offer bargaining game also implements the NBS, even when there are multiple negotiators, under some consistency conditions (Binmore *et al.*, 1986; Krishna and Serrano, 1996). In the Rubinstein bargaining game, a proposer suggests an outcome $\mathbf{x}$, implying that when one x_i is accepted, it is conditional on x_j being accepted. This conditionality is a well-known characteristic of actual international negotiations, as in those leading up to the Kyoto Protocol (as discussed in the Introduction). It is therefore not unreasonable to approximate the outcome of the Kyoto Protocol with the NBS for the participating parties.

However, the procedure adopted in Paris was quite different. As explained in the Introduction, the parties were themselves free to make any NDC they wanted. This change weakened the conditionality aspect that characterised earlier negotiations. Observers have feared that the weight

[11] About bottom-up versus top-down agreements, they write (p. 104): 'These two categories in fact contain several overlapping dimensions, including: "only" domestically legally binding commitments vs internationally legally binding ones (with sanctions in the case of non-compliance); and voluntary commitments (i.e., self-determined contributions) vs commitments based on an explicit formula for sharing the global costs (or "burden") and benefits.' Here, I consider both these dimensions.

on others' payoffs may be less with P&R than with the NBS, because a party has more discretion when deciding on x_i under P&R.

To formalise and isolate this procedural difference, suppose that all parties announce their pledges simultaneously and independently. There is no unilateral commitment, so, if some party rejects $\mathbf{x}$, there will be some delay, Δ, before the parties can revise the pledges. Let $\mathbf{x}^*$ characterise a stationary SPE outcome. If $\mathbf{x}^*$ is expected in the next round, party i approves an arbitrary vector of offers, $\mathbf{x}$, with probability 1 if $U_i(\mathbf{x}) \geq \delta^{\Delta} U_i(\mathbf{x}^*)$ and with probability 0 otherwise.

If δ^{Δ} is known, and $\delta^{\Delta} < 1$, then there is no stationary SPE in which $U_i(\mathbf{x}^*) > 0$ for every i. The proof is by contradiction: if such an SPE existed, party j could deviate by offering slightly less than x_j^*, without fearing that i would reject the offer.

In reality, it is unclear what the opponents will find to be acceptable. With enough uncertainty, it is possible to support a stationary SPE in which $U_i(\mathbf{x}^*) > 0$ for every i. The intuition is that every small deviation from x_j^* may be associated with a small probability that i declines.

To formalise a version of this idea, assume that when i considers whether to reject a vector $\mathbf{x}$, in order to obtain $\mathbf{x}^*$ later, i will apply the discount factor $\delta_{i,t}$, or, equivalently, the discount rate $\theta_{i,t}\rho$, approximated by $\theta_{i,t}\rho = (1 - \delta^{\Delta})/\Delta$. Let every $\theta_{i,t}$ be i.i.d. over time, distributed according to the pdf $f(\cdot)$ on support $[0, \overline{\theta}]$ with expected value 1, so that ρ is the mean of $\theta_{i,t}\rho$. Let the $\theta_{i,t}$ be realised and observed by every party after the offers are made at time t but before the parties decide whether to reject or accept $\mathbf{x}$. (The other parts of the model are left unchanged when the uncertainty is based on temporary shocks.) Assume that $n = 2$ and that f is single peaked.

THEOREM 1 (HARSTAD, 2023). *Consider a stationary SPE in which $U_i(\mathbf{x}^*) > 0$ for both parties. Then,*

$$x_i^* \leq \arg\max_{x_i} U_j(x_i, x_j^*) U_j(x_i, x_j^*)^w, \quad \text{where} \quad w = f(0) \leq \tfrac{1}{2}. \tag{13}$$

The proof in the Appendix draws on the companion paper (Harstad, 2023). There, I prove more general results, and that the first inequality in (13) must bind if there are trembles in the offers.

To conclude, bargaining theory supports the intuition that the weight w (on others' payoffs) is smaller (say, $\underline{w}$) under the pledge-under-review procedure used for the Paris Agreement.

4.2. *Self-Enforcement versus Legally Binding Commitments*

So far, it has been assumed that the parties are able to commit to the pledges for T periods. However, given the incentive to free ride, discussed in Section 3, it is reasonable to also be concerned with the temptation to contribute less at the time when other participants are expected to deliver on their promises.

Because decisions are made simultaneously, a party that 'defects' by not contributing will be able to enjoy the benefit from other participants' contributions in that period. A potential caveat, however, may be that other parties will cease to cooperate thereafter.

To illustrate the power of this mechanism, suppose that the parties revert to BAU (i.e., the non-cooperative MPE) forever as soon as one party has defected by contributing less than pledged.[12]

[12] On the one hand, it is possible to sustain as SPEs harsher punishments than the reversion to BAU. With harsher punishments, a treaty would be self-enforcing under a larger set of circumstances than those derived here. On the other hand, if parties could renegotiate punishments then a treaty would be self-enforcing for a smaller set of parameters.

I henceforth restrict attention to $\iota = 1$.[13] The dynamic game in Section 2 is different from a repeated game because past investments influence BAU and thus all future contributions. When party i invests $y_{i,t}$, then i's contribution will increase by $y_{i,t}$ in every future period, even if the parties revert to BAU. As every $y_{i,t}$ is largest at the beginning of the commitment period, the temptation to defect is also largest in the beginning. Then, the payoff if i defects (by not contributing) is as expressed on the left-hand side of the following inequality. This payoff must be smaller than i's equilibrium payoff on the right-hand side:

$$a\left(\sum_{j\neq i} x_j^* + \frac{\delta}{1-\delta}\sum_{j\neq i} y_{j,1}\right) \leq U_i(\mathbf{x}^*). \tag{14}$$

As noted, i's optimal $y_{i,1}$ depends on x_i. To motivate compliance, x_i cannot be too large, given that the contribution cost is convex in x_i. Given the mapping (6) between the equilibrium contributions and the weight placed on opponents' payoffs (w), and the above expressions for $U_i(\mathbf{x}^*)$, (14) can be written as an upper boundary on w. That is, it can be beneficial to defect on equilibrium pledges unless

$$w \leq \widehat{w} \equiv 2 - 2[1 - \delta(l_1 m_1 + l_2 m_2)]\frac{a(1-\delta^T)}{\alpha(1-\delta)}. \tag{15}$$

PROPOSITION 6. *In a self-enforcing treaty, non-compliance is beneficial unless the contributions are limited or, equivalently, the associated w is smaller than $\widehat{w}$, given by (15). This condition is the same whether n is exogenous or endogenous.*

Interestingly, n drops out from inequality (15), and thus n does not influence whether the bargaining outcome will be self-enforcing. It follows that the incentive constraint (15) is the same whether n is exogenous or endogenous. When n is endogenous, the intuition is that a smaller w motivates a larger n and that, in turn, implies that it is more important for each party that cooperation continue. Technically, the invariance follows because both the cost of the individual contribution and the benefit from the others' contributions are proportional to $(n-1)^2$.

The proof in the Appendix permits the punishment to last for any length $L \leq \infty$ of periods and to be triggered by any probability $\phi \in (0, 1]$. The result holds, qualitatively, for every $L > 0$ and $\phi > 0$. Furthermore, $\widehat{w}$ is shown to increase in both L and ϕ.

If it is difficult to motivate compliance, the parties must find additional ways of raising the cost of non-compliance. In reality, there are several ways of increasing these costs, as the exact wording in an international treaty influences the political and reputational costs if one later defects. Although there exists no world government ready to enforce contracts, it is not irrelevant whether a treaty is called 'legally binding'. IPCC (2014, p. 1020) explains that 'a more legally binding commitment ... signals a greater seriousness by states ... These factors increase the costs of violation (through enforcement and sanctions at international and domestic scales, the loss of mutual cooperation by others, and the loss of reputation and credibility in future negotiations).'

The pledges are not legally binding under the Paris Agreement, but: 'the Kyoto Protocol represents a much harder, more prescriptive approach, including legally binding, quantified emissions limitation targets' (Bodansky and Rajamani, 2018, p. 32). An interpretation of this difference is

[13] A finitely long agreement cannot be self-enforcing if $\iota = 0$, i.e., if one returns to the BAU outcome after period T. In such a situation, there would be no incentive to comply in period T, and thus not in period $T-1$, etc. This observation can rationalise why the Paris Agreement specifies that new pledges must be set every five years (i.e., $\iota = 1$) and is also why I henceforth restrict attention to $\iota = 1$.

that, under the Paris Agreement, the most ambitious equilibrium pledge—represented by $\widehat{w}$—is smaller than the largest possible w under the Kyoto Protocol. In other words, the pledges under the Paris Agreement, when represented by w, must be smaller than the w that can be supported by the legally binding cuts under the Kyoto Protocol.

4.3. *Institutional Complementarities*

According to the reasoning above, Kyoto's conditionality and bindingness are strategic complements. With the NBS, which is associated with a larger w than in my formalisation of P&R, the enforcement capacity must be strengthened. In contrast, if w is small because of P&R, it might be unnecessary to introduce tough penalties on non-compliance. Conversely, if emission cuts must be self-enforcing, because they are not legally binding, then the P&R procedure might be sufficient and the top-down bargaining procedure (or the NBS) is unnecessary.

5. Robustness

This relatively technical section explains (and the Appendix proves) that the propositions hold even if the parties negotiate investment levels or emission taxes (or both) instead of (or in addition to) the x_i. The x_i can also be time dependent, and the investment levels might be decided by profit-maximising firms, without changing the propositions. The results are also quite robust to changes in timing.

(*i*) *Pledging to invest*. Some of the NDCs in the Paris Agreement specify national targets for renewable energy.[14] This possibility can be captured by letting parties decide on the $y_{i,t}$ instead of on the x_i. As discussed in the Appendix, it is straightforward to analyse this scenario: when the $y_{i,t}$, but not the $x_{i,t}$, are pinned down, then i's choice of $x_{i,t}$ will satisfy $b(x_{i,t} - Y_{i,t}) = 0$, just as in BAU, where

$$Y_{i,t+1} \equiv Y_{i,t} + y_{i,t} \quad \text{and} \quad Y_{i,1} \equiv 0.$$

If the investment pledge must be time independent (y_i) throughout a commitment period, then i's continuation value can be written as in Lemma 2, where x_i is replaced by y_i, although the definitions of α and β will be different. The proofs of Propositions 2–6 are thus similar to earlier proofs.

In fact, i's continuation value will be separable in the $\mathbf{y}_t$, where $\mathbf{y}_t = (y_{1,t}, \ldots, y_{n,t})$. Consequently, we can apply (4) when parties negotiate $\mathbf{y}_t$, while keeping fixed the investment levels for other periods. Equation (4) will imply that the P&R outcome for $y_{i,t}$ is

$$y^*_{i,t} = (n-1)w\frac{\delta a/c}{1-\delta}. \tag{16}$$

As this $y_{i,t}$ is time independent, there is no loss for the parties if they restrict attention to time-independent investment levels. For these reasons, the length of the commitment period (T) will not influence payoffs, and any T is here equally good, regardless of the levels of n and w.

(*ii*) *Pledging on emission taxes*. It is also straightforward to allow parties to pledge on domestic emission taxes, instead of on emission cuts. With an emission tax $z_{i,t}$, it is natural that consumption

[14] For example, China pledges to increase the share of non-fossil fuels in its primary energy consumption to around 20%, while India pledges to produce about 40% of its electric power from non–fossil-fuel-based energy resources by 2030. For a recent overview, see http://cait.wri.org/indc/#/.

of fossil fuel be given by the condition in which the marginal benefit of consuming (or the marginal cost of abating) equals the tax: $b(x_{i,t} - Y_{i,t}) = z_{i,t}$. When parties are free to decide their investment levels, they will invest just as in BAU, so $y_{i,t} = 0$. If the emission tax level must be time independent (z_i) throughout the commitment period then i's continuation value can be written as in Lemma 1, where x_i is replaced by z_i, although the definitions of α and β differ. Again, the proofs of Propositions 2–6 are similar to earlier proofs.

In fact, i's continuation value will be separable in the $\boldsymbol{z}_t$, where $\boldsymbol{z}_t = (z_{1,t}, \ldots, z_{n,t})$. Consequently, we can apply (4) when parties negotiate $\boldsymbol{z}_t$, while keeping fixed the emission taxes for other periods. Equation (4) will imply that the P&R outcome for $z_{i,t}$ is

$$z^*_{i,t} = (n-1)wa. \tag{17}$$

As this $z_{i,t}$ is time independent, there is no loss for parties if they restrict attention to time-independent emission taxes. For these reasons, the length of the commitment period (T) will not influence payoffs, and any T is equally good, regardless of the levels of n and w.

As a side remark, it is worth noting that the choice of instrument (i.e., whether parties should negotiate the x_i, y_i or z_i) is also independent of n and w. As proven in the Appendix, negotiating investment levels is better for all parties than negotiating emission taxes if and only if investments are inexpensive and the future is important[15]:

$$\frac{1}{\delta} < 1 + \sqrt{\frac{b}{c}}.$$

(*iii*) *Pledging on investment levels and emission taxes*. It can be shown that party i's continuation value is separable in the $\mathbf{y}_t$ and the $\boldsymbol{z}_t$. Thus, (4) can be applied for each instrument separately, while keeping the other fixed. With this procedure, the outcome is given by the combination of (16) and (17). In this case, we have a 'complete contract' because given the negotiated investment levels (and thus the $Y_{i,t}$), the emission taxes pin down the contribution levels. This game is essentially what we obtain if investments are not part of the game.

(*iv*) *Pledging on investment levels and contribution levels*. Once the investment levels (and thus the $Y_{i,t}$) are pinned down, negotiating $z_{i,t} = b(x_{i,t} - Y_{i,t})$ is equivalent to negotiating $x_{i,t}$. Thus, scenario (*iii*) leads to the same outcome as that which occurs when parties can negotiate every investment level and every contribution level. As in (*i*), and as in the game without investments, the choice of T would be irrelevant, regardless of the n and w levels.[16]

(*v*) *Time-dependent contribution levels*. In scenario (*iv*), one best choice of T is $T = \infty$. With $T = \infty$, it is actually irrelevant that parties have negotiated investment levels in addition to contribution levels. The irrelevance follows because, once the $x_{i,t}$ are given for every time, there is no externality associated with the $y_{i,t}$ and, hence, every party will have incentives to invest optimally, without any need to negotiate $y_{i,t}$. As is shown in the Appendix, the equilibrium time-dependent contribution level is

$$x^*_{i,t} = (n-1)w\frac{a}{b} + (n-1)w\frac{a}{c}\frac{\delta}{1-\delta}t.$$

[15] The comparison to the situation in which the x_i are negotiated is more complex, however.

[16] If parties can negotiate time-independent x_j and y_j, which must stay constant throughout the commitment period, then the parties would strictly prefer $T = 1$. With $T = 1$, the outcome will be the same as with time-dependent policies (scenario (*iv*) and scenario (*iii*)), while $T > 1$ would be less efficient. In contrast to the discussion on the optimal T, in Subsection 3.5, there is no need to have a large T when the first-period investment level can be negotiated, as agreeing on $y_{i,1}$ circumvents the hold-up problem.

Given this pledge, party i prefers to invest as in (16), ensuring that the marginal benefit from consuming (and from cutting emissions) is $b(x^*_{i,t} - ty^*_{i,t}) = (n-1)wa$, which coincides with $z^*_{i,t}$ in (17).

In this situation, it is clear that parties are strictly better off with $T = \infty$ than with $T < \infty$, because with any finite T, the equilibrium $y_{i,t}$ is lower (and less efficient) than the $y_{i,t}$ that would follow in scenario (*iv*), which coincides with the equilibrium $y_{i,t}$ when $T = \infty$. When referring to the trade-off discussed in Subsection 3.5, there is here no reason to reduce T in order to update the pledges when the pledges can be time dependent. It is thus optimal with $T = \infty$ to mitigate the hold-up problem.

Of course, $T < \infty$ continues to be optimal if we introduce, e.g., uncertainty (see below).

(*vi*) *Firms invest*. All three scenarios (*iii*)–(*v*) implement the complete contract outcome, i.e., as when all $y_{i,t}$ and $x_{i,t}$ are negotiated according to P&R. The same outcome can be achieved if parties negotiate $x_{i,t}$ at time t, for $T = 1$, while letting firms invest. The equilibrium pledge $x_{i,t}$ will satisfy $b(x_{i,t} - Y_{i,t}) = (n-1)wa$, which thus also characterises the marginal willingness to pay for another unit of $Y_{i,t}$ at time t. Consequently, the present discounted value of a unit invested today is $\delta(n-1)wa/(1-\delta)$, while the marginal investment cost is $cy_{i,t}$. The two are equalised when profit-maximising price-taking firms decide on $y_{i,t}$ and, then, the result is (16), just as when the parties negotiate the investment levels directly. In this situation, it is clear that parties are strictly better off with $T = 1$ than with $T > 1$ (unless the contribution levels are time dependent). Firms, unlike governments, are not discouraged by the nations' hold-up problem when new pledges are negotiated.[17]

(*vii*) *The timing of* T. Proposition 5 showed that every participant agreed on the choice of T and that this choice was independent of n and w. Thus, the choice of T remains the same whether participants decide on T after the participation stage, before the bargaining-choice stage or in between the two. The timing of T influences neither the equilibrium level of n nor the preference regarding w.[18]

(*viii*) *Multiple participation stages*. Propositions 1–4 continue to hold if there is a participation stage before pledges are negotiated at the beginning of every commitment period (i.e., every T period). In Subsection 3.3, (8) implied the same condition for n, whether or not new commitments would be made after T periods (i.e., whether or not $\iota = 1$). Suppose now that there is a new participation stage after T periods. In an MPE, the identity of the n participants must be the same in every commitment period. Thus, a participant prefers to participate if and only if

$$U_i(\mathbf{x}^*) = \frac{\alpha^2(n-1)^2}{\beta(1-\delta^T)} w\left(1 - \frac{w}{2}\right) \geq \frac{\alpha^2(n-1)(n-2)}{\beta} w + \delta^T U_i(\mathbf{x}^*) \quad \Longrightarrow \quad n \leq 1 + \frac{2}{w},$$

implying that Proposition 3 holds. (Propositions 1 and 2 hold because they take n as given.) Taking this inequality as given, the proofs of Propositions 1–4 remain unchanged.

(*ix*) *Multiple bargaining-choice stages*. Propositions 1–4 also hold if w, as well as n, are endogenously chosen at the beginning of every commitment period, for the same reasons as

[17] If each government can subsidise/tax the firms' investments, it can implement its preferred choice of investment, as described in the previous sections. Then, even the exact equations in Sections 2–3 stay unchanged, one can argue.

[18] However, if T is decided on after the *participation stage* then a suboptimal small T might be preferred if n turns out to be small. The cost of the suboptimally small T can motivate more countries to participate, even if w is large (as in Battaglini and Harstad, 2016). Therefore, with this timing, the trade-off between the procedures is different (Eichner and Schopf, 2022).

in scenario (*viii*). In fact, if some parameters (such as $\underline{n}$ and/or $\overline{n}$) change every T period then Propositions 1–3 characterise the outcome, and Proposition 4 characterises the best bargaining procedure, for every commitment period, regardless of the parameter values after period T. This generalisation implies that Proposition 4 can indeed rationalise a change from one procedure to another, if $\underline{n}$ and/or $\overline{n}$ has changed.[19]

(*x*) *Limited punishments*. When the self-enforcement constraint was discussed, Proposition 6 relied on the assumption that if one party defected then all parties would play BAU forever after. On the one hand, one may argue that it is more realistic to assume that a defection can be observed with probability $\phi < 1$. On the other hand, one may also argue that if cooperation has broken down then parties might renegotiate to start cooperating again. To capture these concerns to some extent, the proof of Proposition 6 permits defection to be punished with a reversion to BAU for $L \leq \infty$ periods with probability $\phi \leq 1$ (while, with probability $1 - \phi$, there is no punishment). The incentive constraint becomes

$$w \leq 2 - 2\left[\frac{1 - \delta(l_1 m_1 + l_2 m_2)}{(1-\delta)(1 - \delta([1 - \phi + \phi\delta^L]))}\right]\frac{a(1-\delta^T)}{\alpha}.$$

A smaller ϕ or L strengthens the incentive constraint, but note that Proposition 6 holds for all $\phi \in (0, 1]$ and $L > 0$.

These generalisations can be summarised in the following proposition (proven in the Appendix).

PROPOSITION 7. *Propositions 2–6 continue to hold if parties pledge-and-review bargain*

(*i*) *investment levels instead of* $\mathbf{x}$*;*
(*ii*) *emission taxes instead of* $\mathbf{x}$*;*
(*iii*) *investment levels and emission taxes instead of* $\mathbf{x}$*;*
(*iv*) *investment levels and* $\mathbf{x}$ *instead of only* $\mathbf{x}$*;*
(*v*) *a time profile* $\{\mathbf{x}_t\}_{t=1}^{\infty}$ *instead of a time-independent* $\mathbf{x}$*;*
(*vi*) $\mathbf{x}$*, while profit-maximising price-taking firms invest;*
(*vii*) T *after the n stage, or before n but after the w stage.*

Furthermore,

(*viii*) *Propositions 1–4 continue to hold if there is a participation stage before every commitment period;*
(*ix*) *Propositions 1–4 continue to hold if both w and n are decided on every commitment period;*
(*x*) *Proposition 6 holds if defection leads to BAU for* $L \in (0, \infty]$ *periods with probability* $\phi \in (0, 1]$.

The optimal level of T varies across the scenarios, but, for every scenario, the optimal T is independent of the bargaining procedure. Obviously, the optimal T, as well as the other results, may depend on many things that are outside of this model, such as policy makers' ability to commit or predict the optimal level of contributions in the distant future. Propositions 1–4 and 6 have thus been derived for any fixed T, and they hold for every T.

[19] The analysis would have been more complicated, however, if parameters also changed within commitment periods.

6. Predictions and Evidence

The results derived in this paper amount to a number of empirical predictions. In Section 4, I argued that the pledge-and-review bargaining procedure, associated with the Paris Agreement, is quite different from the procedure used to negotiate the emission cuts under the Kyoto Protocol. The argument suggested that the weights placed on other parties' payoffs are smaller under the Paris Agreement's P&R procedure (i.e., $\underline{w} < 1$) than they were under Kyoto's 'top-down' procedure with legally binding cuts (then, $\overline{w} > \underline{w}$). With this point of departure, the analysis resulted in five predictions that, according to Section 5, were especially robust. In this section, I summarise the main predictions and relate them to empirical observations.

6.1. *Participation*

Proposition 3 predicts that participation is larger with the P&R procedure associated with the Paris Agreement. As is well known, only 37 countries promised emission cuts for the Kyoto Protocol's first commitment period, but 195 countries have pledged to contribute to the Paris Agreement.

6.2. *Pledges*

Propositions 1 and 2 predict that contributions and investments are smaller when w is small—for any fixed participation level. Proposition 2 is consistent with the criticism mentioned in the introduction as well as with negative experimental evidence on P&R (Barrett and Dannenberg, 2016).

When participation is endogenous, however, every participant's contribution level and investment level is the same for every $w \in W$ in this model. The reason is that x_i increases in both w and n, and equilibrium n falls in w. The prediction that x_i continues to be large even if w is small is consistent with the fact that the pledges in Table 1 are substantial, despite the P&R procedure.

6.3. *From Kyoto to Paris*

Proposition 4 predicts that every participant prefers P&R if and only if $\underline{n}$ is small and/or $\overline{n}$ is large. One may argue that both these developments (i.e., a larger $\overline{n}$ and/or a smaller $\underline{n}$) are in line with changes in world politics over the last couple of decades. In the 1990s, there were a large number of developing countries that could not be expected to contribute much to a global climate policy. Over the next 20 years, some of these became emerging economies that potentially had important roles to play. The number of relevant potential parties, $\overline{n}$, has therefore increased. When $\overline{n}$ is large, it becomes more important to select a procedure that is acceptable even when the set of participants is large.

In fact, already in 2010, Falkner *et al.* (2010, p. 255) wrote: 'the question of how to include the rapidly emerging emitters from the developing world in future mitigation efforts was left unresolved. It was to resurface as a critical stumbling block in the run-up to the 2009 Copenhagen conference'. (The US ratification requirement is referred to as a second hurdle for a Kyoto-style agreement.)

After the first commitment period, seven of the original Annex I countries, which initially signed the Kyoto Protocol, announced that they would not contribute to the Kyoto Protocol's

second commitment period.[20] These withdrawals may be interpreted as a reduction in the number of committed countries, $\underline{n}$.

Either (or both) of these developments makes P&R relatively more attractive for every participant, according to the theory. Thus, the model is consistent with the fact that the parties preferred the Kyoto Protocol in the 1990s, but P&R in the 2010s.

As noted in Subsection 3.4, the theory also predicts that countries that did not contribute to Kyoto should prefer to switch to P&R only under condition (12), which was stronger than condition (11). Thus, it is possible that (11) holds, but (12) fails. In that situation, there will be a disagreement between the North and the South regarding what procedure to choose. In reality, such a disagreement did indeed exist. Bodansky *et al.* (2017, p. 202) wrote: 'Developing countries, for which the Kyoto model has obvious attractions because they are exempt from emissions targets, were keen to extend the protocol for a second and future commitment periods. Kyoto Annex B parties, in contrast, were reluctant to do so, for some countries because of Kyoto's prescriptive architecture, and for others because they did not want to be subject to emissions targets if the US, China, and other large emitters were not.'

6.4. *Commitment Period Length*

Proposition 5 predicts that the optimal commitment period length should be independent from both the bargaining procedure and the coalition size. Thus, if a five-year commitment period was optimal under the Kyoto Protocol, it should be also optimal for the Paris Agreement, according to this result.

Given the many differences between the Kyoto Protocol and the Paris Agreement, it does appear surprising that the two are surprisingly similar regarding how frequently commitments must be updated. Pledges under the Paris Agreement must be updated every five years, and the Kyoto Protocol's first commitment period was also five years (2007–12). It is reasonable that Kyoto's second commitment period would also have been five years, if the parties had not anticipated that a new global treaty would be effective from 2020.[21]

Not only this result but also the mechanism driving it seems to match well with reality. The first argument of the OECD (2018, p. 5) for a five- rather than a ten-year commitment period is: 'More regular opportunities to make technical and fundamental adjustments to NDCs as well as to incorporate effects of technology ....'

6.5. *Legal Status and Compliance*

Returning to Section 4, Proposition 6 shows that the level of w and the strength of enforcement are strategic complements: if the bargaining procedure is the NBS, rather than P&R, the necessary punishment (measured by duration or probability) is more demanding. Hence, the prediction is that these two institutional details may be bundled.

[20] According to the IPCC (2014, p. 1025), 'a number of Annex I countries (Belarus, Canada, Japan, New Zealand, Russia, the United States, and Ukraine) decided not to participate in the second commitment period'.

[21] Bang *et al.* (2016, p. 216) wrote: 'Kyoto, too, aimed for a series of 5-year periods with new and more ambitious commitments in every period.' According to Bodansky *et al.* (2017, p. 203), in 2011, 'Parties disagreed on several issues including: the length of the commitment period—whether it should be five years (like the first commitment period) or eight years (to coincide with the scheduled launch of the 2015 agreement).' In 2012, 'the eight-year duration of the second commitment period was chosen so as to end when the Paris Agreement's NDCs were expected to take effect, and thus to avoid a commitment gap' (p. 205).

In other words, when the Paris Agreement applies P&R bargaining, where w is smaller, it is possible that the incentive constraint holds for this agreement without making it legally binding. In this case, the parties would strictly prefer non-binding commitments if there were tiny costs associated with legally bindingness (e.g., the observed emission level might be only partly under the government's control, etc.). In line with this prediction, the pledges under Paris are not only determined by P&R but are also not legally binding, in contrast to the emission cuts under the Kyoto Protocol.

Of course, when one can raise the cost of non-compliance by modifying the legal status of the agreement, then countries will comply on the equilibrium path regardless of the bargaining procedure. In line with this prediction, the remaining 'thirty-six Kyoto parties [after Canada pulled out] were in full compliance with their first commitment period targets' (Bodansky and Rajamani, 2018, p. 42).

7. Limitations and Future Research

In this paper, I have shown that different bargaining procedures result in several theoretical predictions that are consistent with the differences between the Kyoto Protocol and the Paris Agreement, as well as the transitioning from the former to the latter.

Empirically, the consistency between the predictions and the facts does not prove a causal relationship. The world is changing fast and violating the assumptions imposed in my stylised model. I have allowed for changes in the technology, the investment costs and the need for energy, but other model parameters have been independent of time. I have also abstracted from political economy forces, even though the need for treaty ratification in the United States may have been an important motivation behind the transitioning to the Paris-style agreement. I have also abstracted from technological spillovers, financial transfers and several sources for heterogeneity. The assumptions should be relaxed in future research.[22]

Although the paper focuses on a positive analysis, the reader may instinctively search for normative lessons. One lesson is that pledge and review might not be as inadequate as it at first appears to be; it can actually be preferable to the alternative when participation is endogenous. However, if participation can be encouraged by other means then a more demanding conditional-offer bargaining game becomes preferable. Consequently, the benefit of offering the participants 'club benefits' (such as the lower tariffs) is not, necessarily, that participation will increase (as in Nordhaus, 2015). Instead, the benefit may be that parties can choose a more ambitious bargaining procedure without fearing that participation will fall by too much. By analysing this and related extensions, future research can help by deepening our understanding of the best possible bargaining procedure.

[22] It is possible to extend the model in many other directions as well. Consider, for example, the uncertainty in Gerlagh and Liski (2018). In Harstad (2016), relying on the NBS, both pollution and shocks on the marginal environmental harm accumulate over time. The shocks make it hard to predict optimal pledges and they motivate a small T, while the hold-up problem motivates a large T, especially when there are large technological spillovers. Acemoglu *et al.* (2012) permitted investments in dirty as well as in green technology, Dutta and Radner (2020) transferred from the North to the South, Karp (2017) altruism, Borrero and Rubio (2022) adaptation and Martimort and Sand-Zantman (2016) a mechanism-design approach.

Appendix

I start by reformulating the optimal control problem described in Section 2.

LEMMA 3. *Given the actual pledges,* $\mathbf{x}$*, and the future equilibrium pledges,* $\mathbf{x}^*$*, party* i*'s continuation value is* $V_{i,1}(\mathbf{x}) = V_{i,1}^{BAU} + U_i(\mathbf{x})$*, where*

$$U_i(\mathbf{x}) \equiv \max_{\{y_{i,t}\}_{t=1}^{T}} \sum_{t=1}^{T} \delta^{t-1}\left[a\sum_{j\neq i} x_j - \frac{b}{2}(x_i - Y_{i,t})^2 - \frac{c}{2}y_{i,t}^2\right] + a\frac{\delta^T}{1-\delta}\sum_{j\neq i} Y_{j,T+1} + \delta^T U_i(\mathbf{x}^*),$$

$$Y_{i,t+1} \equiv Y_{i,t} + y_{i,t} \quad \textit{and} \quad Y_{i,1} \equiv 0.$$

The lemma permits the current pledges ($\mathbf{x}$) to be different from those expected in equilibrium in the subsequent commitment period (i.e., $\mathbf{x}^*$). Conveniently, the heterogeneous bliss points and initial technology levels drop out when utility is measured relative to BAU. It is also convenient that the investments' effects on $Y_{j,T+1}$ are captured in terms that do not interact with the future continuation value, $\delta^T U_i(\mathbf{x}^*)$. The additional investments affect the future $V_{i,1}^{BAU}$ but not $U_i(\mathbf{x})$.

Proof of Lemma 3. I will first derive $V_{i,t}^{BAU}$. When we substitute in for $u_{i,t}$, $q_{i,t}^{BAU}$ and $r_{i,t}^{BAU}$ into $V_{i,t}^{BAU} = \sum_{\tau=t}^{\infty} \delta^{\tau-t} u_{i,\tau}$, we can rewrite $V_{i,t}^{BAU}$ as

$$\begin{aligned}
V_{i,t}^{BAU} &= \sum_{\tau=t}^{\infty} \delta^{\tau-t}\left[a\sum_{j\in N}\left(R_{j,\tau} + \frac{a}{b}\right) - \frac{b}{2}\left(\frac{a}{b}\right)^2 - \frac{c}{2}\left(\frac{\delta}{1-\delta}\frac{a}{c}\right)^2\right] \\
&= a\sum_{\tau=t}^{\infty}\sum_{j\in N}\delta^{\tau-t}R_{j,t} + \frac{1}{1-\delta}\left[n\frac{a^2}{b} - \frac{b}{2}\left(\frac{a}{b}\right)^2 - \frac{c}{2}\left(\frac{\delta}{1-\delta}\frac{a}{c}\right)^2\right] \\
&= a\sum_{j\in N}\frac{1}{1-\delta}R_{j,t} + a\sum_{\tau=t}^{\infty}\sum_{j\in N}\frac{\delta^{\tau+1-t}}{1-\delta}r_{i,\tau} + \frac{1}{1-\delta}\left[\frac{a^2}{b}\left(n-\frac{1}{2}\right) - \frac{c}{2}\left(\frac{\delta}{1-\delta}\frac{a}{c}\right)^2\right] \\
&= a\sum_{j\in N}\frac{1}{1-\delta}R_{j,t} + a\sum_{\tau=t}^{\infty}\sum_{j\in N}\frac{\delta^{\tau+1-t}}{1-\delta}\left(\frac{\delta}{1-\delta}\frac{a}{c} - \underline{c}_{i,\tau}\right) \\
&\quad + \frac{a^2}{1-\delta}\left[\frac{1}{b}\left(n-\frac{1}{2}\right) - \frac{1}{2c}\left(\frac{\delta}{1-\delta}\right)^2\right] \\
&= a\sum_{j\in N}\frac{R_{j,t} - \sum_{\tau=t}^{\infty}\delta^{\tau+1-t}\underline{c}_{i,\tau}}{1-\delta} + a\frac{n\delta}{(1-\delta)^2}\left(\frac{\delta}{1-\delta}\frac{a}{c}\right) \\
&\quad + \frac{a^2}{1-\delta}\left[\frac{1}{b}\left(n-\frac{1}{2}\right) - \frac{1}{2c}\left(\frac{\delta}{1-\delta}\right)^2\right] \\
&= a\sum_{j\in N}\frac{R_{j,t} - \sum_{\tau=t}^{\infty}\delta^{\tau+1-t}\underline{c}_{i,\tau}}{1-\delta} + \frac{a^2}{1-\delta}\left[\frac{1}{b}\left(n-\frac{1}{2}\right) + \frac{n\delta^2}{c(1-\delta)^2} - \frac{1}{2c}\left(\frac{\delta}{1-\delta}\right)^2\right] \\
&= a\sum_{j\in N}\frac{R_{j,t} - \sum_{\tau=t}^{\infty}\delta^{\tau+1-t}\underline{c}_{i,\tau}}{1-\delta} + \frac{a^2}{1-\delta}\left(\frac{1}{b} + \frac{1}{c}\left[\frac{\delta}{1-\delta}\right]^2\right)\left(n-\frac{1}{2}\right).
\end{aligned}$$

Similarly, the BAU payoff at time $T+1$ can be written as

$$V_{i,T+1}^{BAU} = \frac{a}{1-\delta}\sum_{j\in N}\left(R_{j,T+1}^{BAU} - \sum_{\tau=T+1}^{\infty}\delta^{\tau-T}\underline{c}_{i,\tau} + Y_{j,T+1}\right) + \frac{a^2}{1-\delta}\left(\frac{1}{b} + \frac{1}{c}\left[\frac{\delta}{1-\delta}\right]^2\right)\left(n-\frac{1}{2}\right),$$

where $Y_{i,T+1}$ measures the additional investments, relative to BAU, thanks to the first commitment period. Each party's present-discounted value of $Y_{i,T+1}$ is $a[\delta^T/(1-\delta)]\sum_j Y_{i,T+1}$, when evaluated in period 1. This term should be added when we derive the additional utility, relative to BAU, when the n parties commit to $\mathbf{x}$ for T periods at time $t=1$ (even if the parties thereafter returned to BAU). The additional utility, relative to BAU, is thus

$$\begin{aligned}
&\sum_{t=1}^{T}\delta^{t-1}\Bigg[a\sum_{j\in N}(q_{j,t}^{BAU}+x_j) - \frac{b}{2}(q_{i,t}^{BAU}+x_i - R_{i,t}^{BAU} - Y_{i,t})^2 \\
&\qquad - \frac{c}{2}\left(\frac{\underline{c}_{i,t}}{c} + r_{i,t}^{BAU} + y_{i,t}\right)^2 - u_{i,t}^{BAU}\Bigg] + a\frac{\delta^T}{1-\delta}\sum_{j\in N}Y_{j,T+1} \\
&= \sum_{t=1}^{T}\delta^{t-1}\Bigg[a\sum_{j\in N}x_j - \frac{b}{2}\left(x_i - Y_{i,t}\right)^2 - b(q_{i,t}^{BAU} - R_{i,t}^{BAU})(x_i - Y_{i,t}) \\
&\qquad - \frac{c}{2}y_{i,t}^2 - c\left(\frac{\underline{c}_{i,t}}{c} + r_{i,t}^{BAU}\right)y_{i,t}\Bigg] + a\delta^T\frac{\sum_{j\in N}Y_{j,T+1}}{1-\delta} \\
&= \sum_{t=1}^{T}\delta^{t-1}\left[a\sum_{j\in N}x_j - \frac{b}{2}(x_i - Y_{i,t})^2 - a(x_i - Y_{i,t}) - \frac{c}{2}y_{i,t}^2 - \frac{\delta a}{1-\delta}y_{i,t}\right] \\
&\quad + a\delta^T\frac{\sum_{j\in N}Y_{j,T+1}}{1-\delta} \\
&= \sum_{t=1}^{T}\delta^{t-1}\left[a\sum_{j\neq i}x_j - \frac{b}{2}(x_i - Y_{i,t})^2 - \frac{c}{2}y_{i,t}^2 + aY_{i,t} - \delta a\frac{Y_{i,t+1}-Y_{i,t}}{1-\delta}\right] \\
&\quad + a\delta^T\frac{\sum_{j\in N}Y_{j,T+1}}{1-\delta} \qquad \text{(A1)} \\
&= \sum_{t=1}^{T}\delta^{t-1}\left[a\sum_{j\neq i}x_j - \frac{b}{2}(x_i - Y_{i,t})^2 - \frac{c}{2}y_{i,t}^2\right] + a\frac{\delta^T}{1-\delta}\sum_{j\neq i}Y_{j,T+1},
\end{aligned}$$

where the last equality follows because the three terms with $Y_{i,\tau}$ in (A1) sum to zero for each $\tau = \{2,\ldots,T+1\}$ and because $Y_{i,1}=0$.

When the parties do *not* play BAU after the first commitment period then, in order to obtain i's total additional payoff relative to BAU, we must add the additional payoff $\delta^T U_i(\mathbf{x}^*)$, where $U_i(\mathbf{x}^*)$ is the equilibrium additional utility relative to BAU, in order to get $U_i(\mathbf{x})$ in Lemma 3. □

Proof of Proposition 1. Lemma 3 defines an optimal-control problem with control $y_{i,t}$. Note that the terminal value for $Y_{i,T+1}$ is zero because $U_i(\mathbf{x})$ is measured relative to $V_{i,1}^{BAU}$: this implies that $y_{i,T} = 0$, i.e., the investment level in the final period coincides with the equilibrium investment level in BAU. In other words, there is no *additional* investment in the final period.

When λ_t defines the shadow value of the stock $Y_{i,t}$, evaluated at time 1, the discrete-time Hamiltonian can be written as[23]

$$H_t = \delta^{t-1}\left[a\sum_{j\neq i} x_j - \frac{b}{2}(x_i - Y_{i,t})^2 - \frac{c}{2}y_{i,t}^2\right] + \lambda_{i,t+1}(Y_{i,t} + y_{i,t})$$

with first-order conditions

$$y_{i,t} = \arg\max_{y_{i,t}} H_t = \lambda_{i,t+1}/c\delta^{t-1},$$

adjoint equation

$$\lambda_{i,t+1} - \lambda_{i,t} = -\frac{\partial H_t}{Y_{i,t}} = -\delta^{t-1}b(x_i - Y_{i,t})$$

and terminal condition

$$\lambda_{i,T+1} = 0 \quad \Longleftrightarrow \quad y_{i,T} = 0.$$

Combining the first two conditions and (1), we get the second-order difference equation:

$$c\delta^{t-2}(Y_{i,t} - Y_{i,t-1}) - c\delta^{t-1}(Y_{i,t+1} - Y_{i,t}) = \delta^{t-1}(x_i - Y_{i,t})b$$
$$\Longrightarrow \quad -Y_{i,t+1} + \left(\frac{1}{\delta} + 1 + \frac{b}{c}\right)Y_{i,t} - \frac{1}{\delta}Y_{i,t-1} = x_i b/c.$$

This has the solution (see, e.g., Sydsaeter and Hammond, 1995, pp. 751–3)

$$Y_{i,t} = A_1 m_1^{t-1} + A_2 m_2^{t-1} + x_i, \tag{A2}$$

where

$$m_1 = \frac{1}{2}\left(\frac{1}{\delta} + 1 + \frac{b}{c}\right) - \frac{1}{2}\sqrt{\left(\frac{1}{\delta} + 1 + \frac{b}{c}\right)^2 - \frac{4}{\delta}} \in (0, 1),$$
$$m_2 = \frac{1}{2}\left(\frac{1}{\delta} + 1 + \frac{b}{c}\right) + \frac{1}{2}\sqrt{\left(\frac{1}{\delta} + 1 + \frac{b}{c}\right)^2 - \frac{4}{\delta}} > 1.$$

[23] I here apply Pontryagin's maximum principle for discrete time problems. For a general characterisation and proof, see, for example, Leonard and van Long (1992, pp. 129–33).

The constants A_1 and A_2 can be derived from the initial condition $Y_{i,1} = 0$, implying that $A_1 + A_2 = -x_i$, and the terminal condition, $y_{i,T} = 0$, implying that

$$y_{i,T} = Y_{i,T+1} - Y_{i,T} = A_1 m_1^T \left(1 - \frac{1}{m_1}\right) - (A_1 + x_i) m_2^T \left(1 - \frac{1}{m_2}\right) = 0$$

$$\implies \quad A_1 = -\frac{m_2^T(1 - 1/m_2)}{m_1^T(1/m_1 - 1) + m_2^T(1 - 1/m_2)} x_i$$

$$\text{and} \quad A_2 = -A_1 - x_i$$

$$= \frac{m_2^T(1 - 1/m_2)}{m_1^T(1/m_1 - 1) + m_2^T(1 - 1/m_2)} x_i - x_i$$

$$= -\frac{m_1^T(1/m_1 - 1)}{m_1^T(1/m_1 - 1) + m_2^T(1 - 1/m_2)} x_i.$$

With the definitions $l_1 = -A_1 x_i$ and $l_2 = -A_2 x_i$, (A2) can be written as in Proposition 1. □

Proof of Lemma 1. By substituting in for $y_{i,t}$ and $Y_{i,t}$ into $U_{i,1}(\mathbf{x})$, defined in Lemma 3, we get

$$U_i(\mathbf{x}) - \delta^T U_i(\mathbf{x}^*) = \sum_{t=1}^{T} \delta^{t-1} \left[a \sum_{j \neq i} x_j - \frac{b}{2}(x_i - Y_{i,t})^2 - \frac{c}{2} y_{i,t}^2 \right] + a \frac{\delta^T}{1 - \delta} \sum_{j \neq i} Y_{j,T+1}$$

$$= \sum_{t=1}^{T} \delta^{t-1} \left[a \sum_{j \neq i} x_j - \frac{b}{2} x_i^2 (l_1 m_1^{t-1} + l_2 m_2^{t-1})^2 \right.$$

$$\left. - \frac{c}{2} [x_i (l_1 m_1^{t-1}[1 - m_1] - l_2 m_2^{t-1}[m_2 - 1])]^2 \right]$$

$$+ a \frac{\delta^T}{1 - \delta} \sum_{j \neq i} Y_{j,T+1}$$

$$= \alpha \sum_{j \neq i} x_j + \beta x_i^2 / 2$$

if just

$$\alpha \equiv \sum_{t=1}^{T} \delta^{t-1} a + a \frac{\delta^T}{1 - \delta} \frac{Y_{j,T+1}}{x_j} = \frac{a}{1 - \delta} [1 - \delta^T (l_1 m_1^{T-1} + l_2 m_2^{T-1})]$$

$$\text{and} \quad \beta \equiv \sum_{t=1}^{T} \delta^{t-1} [b(l_1 m_1^{t-1} + l_2 m_2^{t-1})^2 + c[(l_1 m_1^{t-1}[1 - m_1] - l_2 m_2^{t-1}[m_2 - 1])]^2].$$

This completes the proof. □

Proof of Propositions 2–5. The proof of Proposition 2 follows from the earlier lemmata, while Propositions 3 and 4 follow from the reasoning in the text. Proposition 5 follows straightforwardly from the equilibrium continuation values, derived above. □

Proof of Theorem 1. The original and more general proof is in Harstad (2023). Here, let $i, j \in \{1, 2\}$, $i \neq j$. Consider i's deviation, leading to $\mathbf{x}^i = (x_i^* + dx_i, x_j^*)$ for small $dx_i < 0$, assumed to increase U_i but decrease U_j. Party $j \neq i$ rejects $\mathbf{x}^i$ if and only if

$$U_j(\mathbf{x}^i) < (1 - \theta_{j,t}\rho\Delta)U_j(\mathbf{x}^*) \quad \Longleftrightarrow \quad \theta_{j,t} < \widetilde{\theta}_j(\mathbf{x}^i) \equiv 0, \frac{U_j(\mathbf{x}^*) - U_j(\mathbf{x}^i)}{\rho\Delta U_j(\mathbf{x}^*)}.$$

When $dx_i \uparrow 0$, the probability that j rejects can be written as

$$P(\mathbf{x}^i) = \Pr(\theta_{j,t} < \widetilde{\theta}_j(\mathbf{x}^i)) \to \frac{\partial U_j(\mathbf{x}^i)/\partial x_i}{\rho\Delta U_j(\mathbf{x}^*)}(-dx_i)f(0). \tag{A3}$$

Consider party i's expected payoff when proposing $x_i^i = x_i^* + dx_i$ once. This payoff, written on the left-hand side in the following inequality, is smaller than i's payoff if i sticks to the equilibrium x_i^* if and only if

$$(1 - P(\mathbf{x}^i))\left(U_i(\mathbf{x}^*) + dx_i\frac{\partial U_i(\mathbf{x}^*)}{\partial x_i}\right) + P(\mathbf{x}^i)(1 - \rho\Delta)U_i(\mathbf{x}^*) \leq U_i(\mathbf{x}^*).$$

When we substitute in with (A3), divide both sides by $|dx_i|$ and $dx_i \uparrow 0$, the inequality can be rewritten as

$$\frac{\partial U_i(\mathbf{x}^*)/\partial x_i}{U_i(\mathbf{x}^*)} \leq \frac{\partial U_j(\mathbf{x}^*)/\partial x_i}{U_j(\mathbf{x}^*)}f(0),$$

which coincides with the f.o.c. of the maximisation problem in (13) given the constraint $x_i \leq x_i^*$ (which (A3) requires). Here, $x_i > x_i^*$ does not raise $P(\mathbf{x}^i)$ and cannot be optimal for i.

Finally, note that if f is single peaked and symmetric with mean 1 then $f(0) \leq f(\epsilon)$ for every $\epsilon \in [0, 1]$, so

$$f(0) = \int_0^1 f(0)d\epsilon \leq \int_0^1 f(\epsilon)d\epsilon = \frac{1}{2}\int_0^2 f(\epsilon)d\epsilon = \frac{1}{2}.$$

This completes the proof. □

Proof of Proposition 6. If i defects by not contributing at time t then i can still benefit $a\sum_{j\neq i} x_j + [a\delta/(1-\delta)]\sum_{j\neq i} y_{j,t}$, because j's investments will raise j's contribution in the future, even when the parties return to BAU. This benefit is largest at $t = 1$, because $y_{j,t}$ is decreasing in $t \in \{1, \ldots, T\}$, as noted already.

When defection is punished by a reversion to BAU for $L \leq \infty$ periods with probability $\phi \in (0, 1]$, then compliance (giving payoff U_i^*) is better at time $t = 1$ if

$$a\sum_{j\neq i} x_j + \frac{a\delta}{1-\delta}\sum_{j\neq i} y_{j,1} + \delta(1 - \phi + \phi\delta^L)U_i^* \leq U_i^*.$$

When we substitute in for $y_{j,1}$, x_j^* and U_i^*, this inequality becomes

$$a\left(\sum_{j\neq i} x_j^* + \frac{\delta}{1-\delta}\sum_{j\neq i} y_{j,1}^*\right) \leq [1-\delta(1-\phi+\phi\delta^L)]U_i^*$$

$$\iff \quad a\left[1+\frac{\delta}{1-\delta}(1-l_1m_1-l_2m_2)\right]\sum_{j\neq i} x_j^* \leq [1-\delta(1-\phi+\phi\delta^L)]\frac{\alpha^2(n-1)^2}{\beta(1-\delta^T)}w\left(1-\frac{w}{2}\right)$$

$$\iff \quad \frac{a(1-\delta^T)}{\alpha[1-\delta(1-\phi+\phi\delta^L)]}\left[\frac{1-\delta(l_1m_1+l_2m_2)}{1-\delta}\right] \leq 1-\frac{w}{2}$$

$$\iff \quad w \leq 2-2\frac{1-\delta(l_1m_1+l_2m_2)}{(1-\delta)[1-\delta(1-\phi+\phi\delta^L)]}\frac{a(1-\delta^T)}{\alpha},$$

which equals (15) when $\phi = 1$ and $L \to \infty$. □

Proof of Proposition 7. (*i*) *Contracts on investments*. I will first permit the negotiated $\mathbf{y}_t = (y_{1,t}, \ldots, y_{n,t})$ to be time dependent, so that $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_T)$ is a matrix. Lemma 1 presents a reformulation of the problem and (when we remove the max-operator) it holds regardless of how the $x_{i,t}$ and the $y_{i,t}$ are decided on. When $y_{i,t}$ is committed to, but not $x_{i,t}$, the latter follows straightforwardly from i's maximisation problem and, just as in BAU,

$$q_{i,t} - R_{i,t} = a/b \quad \Longrightarrow \quad x_i = Y_{i,t}.$$

The continuation value can thus be written as a function of the investments matrix $\mathbf{y}$:

$$U_i(\mathbf{y}) = \sum_{t=1}^{T}\delta^{t-1}\left[a\sum_{j\neq i}\sum_{t'=1}^{t-1} y_{j,t'} - \frac{c}{2}y_{i,t}^2\right] + a\frac{\delta^T}{1-\delta}\sum_{j\neq i}\sum_{t'=1}^{T} y_{j,t'} + \delta^T U_i(\mathbf{y}^*)$$

$$\iff \quad U_i(\mathbf{y}) - \delta^T U_i(\mathbf{y}^*) = \sum_{t=1}^{T}\left[\alpha_t\sum_{j\neq i} y_{j,t} - \frac{\beta_t}{2}y_{i,t}^2\right],$$

where

$$\alpha_t = \frac{a\delta^t}{1-\delta} \quad \text{and} \quad \beta_t = \delta^{t-1}c.$$

If we require a time-independent $y_{j,t} = y_j$, we can write

$$U_i(\mathbf{y}) - \delta^T U_i(\mathbf{y}^*) = \alpha\sum_{j\neq i} y_j - \frac{\beta}{2}y_i^2,$$

where

$$\alpha = \delta a\frac{1-\delta^T}{(1-\delta)^2} \quad \text{and} \quad \beta = \sum_{t=1}^{T}\delta^{t-1}c = c\frac{1-\delta^T}{1-\delta}.$$

Just as before, i's payoff is in the form of (5). Consequently, the proofs for the other propositions follow the same steps as above. Analogously to (4), we get, for example,

$$y_j^* = w(n-1)\alpha/\beta = w(n-1)\frac{\delta a/c}{1-\delta}.$$

Time-dependent investment levels. As i's payoff is separable in the $y_{j,t}$, we can apply (4) for each $\mathbf{y}_t$, if we fix the investment levels for the other periods, in order to get

$$y_{j,t}^* = w\,(n-1)\,\alpha_t/\beta_t = w\,(n-1)\,\frac{\delta a/c}{1-\delta},$$

which equals y_j^*. Hence, the restriction to time-independent investment levels is non-binding: the equilibrium is the same in both cases.

The choice of T is irrelevant in both cases, because the equilibrium continuation value is

$$\begin{aligned} U_i(\mathbf{y}^*) &= \delta a \frac{1}{(1-\delta)^2}(n-1)^2 w \frac{\delta a/c}{1-\delta} - \frac{c/2}{1-\delta}\left[(n-1)w\frac{\delta a/c}{1-\delta}\right]^2 \\ &= \frac{[\delta a(n-1)]^2}{c(1-\delta)^3} w\left(1-\frac{w}{2}\right). \end{aligned}$$

(*ii*) *Contracts on carbon tax*. I will first permit $\mathbf{z}_t = (z_{1,t}, \ldots, z_{n,t})$ to be time dependent, so that $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_T)$ is a matrix.

With an emission tax equal to $z_{i,t}$, collected by the government in country i, the equilibrium ensures that the marginal benefit when consuming fossil fuel (or the marginal abatement cost) equals $z_{i,t}$. This implies that

$$x_{i,t} - Y_{i,t} = z_{i,t}/b,$$

and, therefore, i's continuation value can be written as the function

$$U_i(\mathbf{z}) = \sum_{t=1}^{T} \delta^{t-1}\left[a\sum_{j\neq i}(z_{j,t}/b + Y_{j,t}) - \frac{z_{i,t}^2}{2b} - \frac{c}{2}y_{i,t}^2\right] + a\frac{\delta^T}{1-\delta}\sum_{j\neq i} Y_{j,T+1} + \delta^T U_i(\mathbf{z}^*),$$

so there is no value for i to invest beyond the BAU levels, and $y_{i,t} = 0$, so

$$U_i(\mathbf{z}) - \delta^T U_i(\mathbf{z}^*) \equiv \sum_{t=1}^{T}\delta^{t-1}\left[a\sum_{j\neq i} z_{j,t}/b - \frac{z_{i,t}^2}{2b}\right] = \sum_{t=1}^{T}\left[\alpha_t \sum_{j\neq i} z_{j,t} - \frac{\beta_t}{2}z_{i,t}^2\right],$$

where

$$\alpha_t = a\delta^{t-1}/b \quad \text{and} \quad \beta_t = \delta^{t-1}/b.$$

If the emission tax is time independent, we can write

$$U_i(\mathbf{z}) - \delta^T U_i(\mathbf{z}^*) = \alpha\sum_{j\neq i} z_j - \frac{\beta}{2}z_i^2,$$

where

$$\alpha = \frac{a}{b}\frac{1-\delta^T}{1-\delta} \quad \text{and} \quad \beta = \frac{1}{b}\frac{1-\delta^T}{1-\delta}.$$

In this case, (4) implies that

$$z_i^* = w(n-1)\alpha/\beta = w(n-1)a.$$

Time-dependent tax. As i's payoff is separable in the $z_{j,t}$, we can apply (4) for each $\mathbf{z}_t$, if we fix the emission tax levels for the other periods, in order to get

$$z_{j,t}^* = w(n-1)\alpha_t/\beta_t = w(n-1)a,$$

which equals z_i^*. Hence, the restriction to time-independent emission tax levels is non-binding: the equilibrium is the same in both cases.

The choice of T is also irrelevant in both cases, because the equilibrium continuation value is

$$U_i(\mathbf{z}^*) = \frac{a}{b}\frac{1}{1-\delta}(n-1)^2 wa - \frac{1}{2}\frac{1}{b}\frac{1}{1-\delta}[(n-1)wa]^2 = \frac{[a(n-1)]^2}{b(1-\delta)}w\left(1-\frac{w}{2}\right).$$

By comparison, a tax gives higher payoff than an investment agreement if

$$\frac{[a(n-1)]^2}{b(1-\delta)} > \frac{[\delta a(n-1)]^2}{c(1-\delta)^3} \iff c(1-\delta)^2 > b\delta^2 \iff \frac{1}{\delta} > 1+\sqrt{\frac{b}{c}}.$$

Clearly, the investment agreement is better if investments are inexpensive and the tax ineffective (because b is large). If δ is large, investments are, in effect, less expensive, and thus the investment agreement is more attractive.

(*iii*) *Combining (i) and (ii).* When the parties face both a matrix of emission taxes and a matrix of investment levels, i's payoff can be written as

$$\begin{aligned} U_i(\mathbf{x}) - \delta^T U_i(\mathbf{x}^*) &\equiv \sum_{t=1}^{T}\delta^{t-1}\left[a\sum_{j\neq i}\left(\frac{z_{j,t}}{b}+Y_{j,t}\right) - \frac{z_{j,t}^2}{2b} - \frac{c}{2}y_{i,t}^2\right] + a\frac{\delta^T}{1-\delta}\sum_{j\neq i}Y_{j,T+1} \\ &= \left[\sum_{t=1}^{T}\delta^{t-1}\left(a\sum_{j\neq i}Y_{j,t} - \frac{c}{2}y_{i,t}^2\right) + a\frac{\delta^T}{1-\delta}\sum_{j\neq i}Y_{j,T+1}\right] \\ &\quad + \left[\sum_{t=1}^{T}\delta^{t-1}\left(a\sum_{j\neq i}\frac{z_{j,t}}{b} - \frac{z_{j,t}^2}{2b}\right)\right], \end{aligned}$$

where the first (second) bracket can be recognised as i's payoff in the situation when only the investment levels (the emission taxes) were negotiated. The two problems are thus separable, and the results above continue to hold when the parties can negotiate both policy instruments. In this case, the additional payoff, relative to BAU, is also the sum of the two additional payoffs, derived above:

$$U_i(\mathbf{y}^*) + U_i(\mathbf{z}^*) = \left[\frac{1}{c(1/\delta-1)^2} + \frac{1}{b}\right]\frac{[a(n-1)]^2}{(1-\delta)}w\left(1-\frac{w}{2}\right).$$

(*iv*) *Complete contracts*. When the parties negotiate the investment levels, the $z_{j,t}$ pin down the $x_{j,t}$, given the $y_{j,t}$, so negotiating the $z_{j,t}$ is then equivalent to negotiating the $x_{j,t}$. Also, when the $y_{j,t}$ and the $x_{j,t}$ are negotiated, the contract is complete and the choice of T is irrelevant. One optimal T is thus $T=\infty$.

(*v*) *Time path for x*. When the $y_{j,t}$ and the $x_{j,t}$ are negotiated, one optimal T is $T=\infty$. In this situation, pinning down the $x_{j,t}$ is equivalent to pinning down both the $y_{j,t}$ and the $x_{j,t}$, because

there is no externality when it comes to the $y_{j,t}$ (given every future $x_{j,t}$) and, hence, every party will invest optimally, without any need to specify the investment levels. This reasoning completes the proof but, to illustrate, consider the time profile for the contribution levels when the parties negotiate both the emission taxes and the investment levels:

$$x_{i,t} = (n-1)wa/b + t(n-1)w\frac{\delta a/c}{1-\delta}.$$

Given this path, optimal investments, from i's point of view, are

$$\begin{aligned} cy_{i,t-1} - \delta cy_{i,t} &= \delta b(x_{i,t} - Y_{i,t}) \\ &= \delta b\left((n-1)wa/b + t(n-1)w\frac{\delta a/c}{1-\delta} - t(n-1)w\frac{\delta a/c}{1-\delta}\right) \\ &= \delta b((n-1)wa/b) \\ \Longrightarrow \quad y_{i,t-1} &= \frac{\delta(n-1)wa}{c(1-\delta)}, \end{aligned}$$

just as in the optimal contract. So, the combination of negotiating investment levels and emission taxes is indeed equivalent to pinning down the path of $x_{i,t}$.

(*vi*) *Firms*. It suffices to prove that, when $T = 1$ and the parties negotiate $x_{i,t}$ at the start of every period t, and the firms invest to maximise profit, then the outcome coincides with the outcome when all the $y_{i,t}$ and the $x_{i,t}$ are negotiated at the very beginning.

When only this period's $x_{i,t}$ are negotiated at the start of period t, then P&R implies that

$$b(x_{i,t} - Y_{i,t}) = aw(n-1).$$

Firms invest such as to equalise the marginal investment cost to the present-discounted value of their investment, where the willingness to pay for more $R_{i,t}$ equals $b(q_{i,t} - R_{i,t})$ at time t. Thus,

$$\begin{aligned} cr_{i,t} + \underline{c}_{i,t} &= \sum_{t=1}^{\infty} \delta^t b(q_{i,t} - R_{i,t}) \\ &= \sum_{t=1}^{\infty} \delta^t b(q_{i,t}^{BAU} - R_{i,t}^{BAU} + x_{i,t} - Y_{i,t}) \\ &= \sum_{t=1}^{\infty} \delta^t b\left(\frac{a}{b} + x_{i,t} - Y_{i,t}\right) \\ &= \sum_{t=1}^{\infty} \delta^t b\left(\frac{a}{b} + \frac{a}{b}w(n-1)\right) \\ &= \frac{\delta}{1-\delta} b\left(\frac{a}{b} + \frac{a}{b}w(n-1)\right). \end{aligned}$$

With

$$r_{i,t} = r_{i,t}^{BAU} + y_{i,t} \quad \text{and} \quad r_{i,t}^{BAU} = \frac{\delta}{1-\delta}\frac{a}{c} - \underline{c}_{i,t},$$

we get $cy_{i,t} = \delta aw(n-1)/(1-\delta)$, as with complete contracts.

Scenarios (*vii*)–(*ix*) are trivial and thus omitted.

(*x*) *Compliance*. In all the above situations and also in the basic model if $c \to \infty$ then U_i^* is independent of T and it can, when the policy instrument is given by the matrix $\boldsymbol{\psi} = (\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_K)$, where $\boldsymbol{\psi}_k = (\boldsymbol{\psi}_{1,k}, \ldots \boldsymbol{\psi}_{n,k})$ for each $k \in \{1, \ldots, K\}$, be written as (for some constants α_k and β_k)

$$U_i^* = \sum_{k\in\{1,\ldots,K\}} \frac{1}{1-\delta}\left[\alpha_k' \sum_{j\neq i} \psi_{j,k} - \frac{\beta_k'}{2}\psi_{j,k}^2\right], \quad \text{so} \quad \psi_{j,k} = w(n-1)\alpha_k'/\beta_k',$$

from (4). If defection is punished by reverting to BAU for L periods with probability ϕ then the incentive constraint is

$$\begin{aligned}
&\sum_{k\in\{1,\ldots,K\}} \alpha_k' \sum_{j\neq i} \psi_{j,k} + \delta(1-\phi+\phi\delta^L)U_i^* \leq U_i^* \\
\iff \quad &\sum_{k\in\{1,\ldots,K\}} \frac{(n-1)^2(\alpha_k')^2}{\beta_k'} w \\
&\qquad \leq \sum_{k\in\{1,\ldots,K\}} \frac{1-\delta(1-\phi+\phi\delta^L)}{1-\delta}\left[\frac{(n-1)^2(\alpha_k')^2}{\beta_k'} w - \frac{\beta_k'}{2}\left[\frac{(n-1)\alpha_k'}{\beta_k'} w\right]^2\right] \\
\iff \quad &1 \leq \frac{1-\delta(1-\phi+\phi\delta^L)}{1-\delta}\left[1 - \frac{1}{2}w\right] \\
\iff \quad &w \leq 2 - 2\frac{1-\delta}{1-\delta(1-\phi+\phi\delta^L)} \\
&\qquad = 2\frac{1-\delta(1-\phi+\phi\delta^L) - 1 + \delta}{1-\delta(1-\phi+\phi\delta^L)} \\
&\qquad = 2\delta\frac{\phi(1-\delta^L)}{1-\delta(1-\phi+\phi\delta^L)},
\end{aligned}$$

which simplifies to $w \leq 2\delta$ if $\phi = 1$ and $L \to \infty$. □

University of Oslo, Norway

References

Acemoglu, D., Aghion, P., Bursztyn, L. and Hemous, D. (2012). 'The environment and directed technical change', *American Economic Review*, vol. 102(1), pp. 131–66.

Aldy, J.E., Barrett, S. and Stavins, R.N. (2003). 'Thirteen plus one: A comparison of global climate policy architectures', *Climate Policy*, vol. 3(4), pp. 373–97.

Bang, G., Hovi, J. and Skodvin, T. (2016). 'The Paris agreement: Short-term and long-term effectiveness', *Politics and Governance*, vol. 4(3), pp. 209–18.

Barrett, S. (1994). 'Self-enforcing international environmental agreements', *Oxford Economic Papers*, vol. 46, pp. 878–94.

Barrett, S. (2002). 'Consensus treaties', *Journal of Institutional and Theoretical Economics*, vol. 158(4), pp. 529–47.

Barrett, S. and Dannenberg, A. (2016). 'An experimental investigation into 'pledge and review' in climate negotiations', *Climatic Change*, vol. 138(1), pp. 339–51.

Battaglini, M. and Harstad, B. (2016). 'Participation and duration of environmental agreements', *Journal of Political Economy*, vol. 124(1), pp. 160–204.

Battaglini, M. and Harstad, B. (2020). 'The political economy of weak treaties', *Journal of Political Economy*, vol. 128(2), pp. 544–90.

Beccherle, J. and Tirole, J. (2011). 'Regional initiatives and the cost of delaying binding climate change agreements', *Journal of Public Economics*, vol. 95, pp. 1339–48.
Bernauer, T., Kalbhenn, A., Koubi, V. and Spilker, G. (2013). 'Is there a "depth versus participation" dilemma in international cooperation?', *Review of International Organization*, vol. 8(4), pp. 477–97.
Binmore, K., Rubinstein, A. and Wolinsky, A. (1986). 'The Nash bargaining solution in economic modelling', *The RAND Journal of Economics*, vol. 17(2), pp. 176–88.
Bloch, F. (2018). 'Coalitions and networks in oligopolies', in (L. Corchon and M. Marini, eds.), *Handbook of Game Theory and Industrial Organization*, pp. 373–91, Cheltenham: Edward Elgar.
Bodansky, D., Brunnee, J. and Rajamani, L. (2017). *International Climate Change Law*, New York: Oxford University Press.
Bodansky, D. and Rajamani, L. (2018). 'The evolution and governance architecture of the United Nations climate change regime', in (U. Luterbacher and D. Sprinz, eds.), *Global Climate Policy: Actors, Concepts, and Enduring Challenges*, pp. 13–66, Cambridge, MA: MIT Press.
Borrero, M. and Rubio, S.J. (2022). 'An adaptation–mitigation game: Does adaptation promote participation in international environmental agreements?', *International Environmental Agreements: Politics, Law and Economics*, vol. 22, pp. 439–79.
Calvo, E. and Rubio, S.J. (2012). 'Dynamic models of international environmental agreements: A differential game approach', *International Review of Environmental and Resource Economics*, vol. 6, pp. 289–339.
Caparrós, A. (2016). 'Bargaining and international environmental agreements', *Environmental and Resource Economics*, vol. 65(1), pp. 5–31.
Caparrós, A. (2020). 'Pledge and implement bargaining in the Paris Agreement on climate change', Working Paper, 2020-03, Instituto de Políticas y Bienes Públicos (IPP) CSIC.
Carraro, C. and Siniscalco, D. (1993). 'Strategies for the international protection of the environment', *Journal of Public Econonomics*, vol. 52(3), pp. 309–28.
d'Aspremont, C., Jacquemin, A., Gabszewicz, J.J. and Weymark, J.A. (1983). 'On the stability of collusive price leadership', *The Canadian Journal of Economics*, vol. 16(1), pp. 17–25.
Dutta, P.K. and Radner, R. (2004). 'Self-enforcing climate-change treaties', *Proceedings of the National Academy of Science*, vol. 101, pp. 4746–51.
Dutta, P.K. and Radner, R. (2006). 'A game-theoretic approach to global warming', *Advances in Mathematical Economics*, vol. 8, pp. 135–53.
Dutta, P.K. and Radner, R. (2020). 'The Paris accord and the green climate fund: A Coase theorem', Mimeo, Columbia University.
Eichner, T. and Schopf, M. (2022). 'Self-enforcing climate agreements: Kyoto versus Paris', Mimeo, University of Hagen.
Falkner, R., Stephan, H. and Vogler, J. (2010). 'International climate policy after Copenhagen: Towards a "building blocks" approach', *Global Policy*, vol. 1(3), pp. 252–62.
Finus, M. and Maus, S. (2008). 'Modesty may pay!', *Journal of Public Economic Theory*, vol. 10, pp. 801–26.
Gerlagh, R. and Liski, M. (2018). 'Carbon prices for the next hundred years', Economic Journal, vol. 128(609), pp. 728–57.
Gilligan, M.J. (2004). 'Is there a broader-deeper trade-off in international multilateral agreements?', *International Organization*, vol. 58(3), pp. 459–84.
Gollier, C. and Tirole, J. (2015). 'Making climate agreements work', *The Economist*, 1 June. https://www.economist.com/free-exchange/2015/06/01/making-climate-agreements-work.
Golosov, M., Hassler, J., Krusell, P. and Tsyvinski, A. (2014). 'Optimal taxes on fossil fuel in general equilibrium', *Econometrica*, vol. 82(1), pp. 41–88.
Harris, M. and Holmstrom, B. (1987). 'On the duration of agreements', *International Economic Review*, vol. 28(2), pp. 389–406.
Harstad, B. (2012). 'Climate contracts: A game of emissions, investments, negotiations, and renegotiations', *Review of Economic Studies*, vol. 79(4), pp. 1527–57.
Harstad, B. (2016). 'The dynamics of climate agreements', *Journal of the European Economic Association*, vol. 14(3), pp. 719–52.
Harstad, B. (2023). 'Pledge-and-review bargaining', *Journal of Economic Theory*, vol. 207, 105574.
Harstad, B., Lancia, F. and Russo, A. (2019). 'Compliance technology and self-enforcing agreements', *Journal of the European Economic Association*, vol. 17(1), pp. 1–30.
Harstad, B., Lancia, F. and Russo, A. (2022). 'Prices vs. quantities for self-enforcing agreements', *Journal of Environmental Economics and Management*, vol. 111, 102595.
Hoel, M. (1992). 'International environmental conventions: The case of uniform reductions of emissions', *Environmental and Resource Economics*, vol. 2(2), pp. 141–59.
IPCC. (2014). *Climate Change 2014: Mitigation of Climate Change. Contribution of Working Group III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, Cambridge: Cambridge University Press.
Karp, L. (2017). 'Provision of a public good with multiple dynasties', Economic Journal, vol. 127(607), pp. 2641–64.
Karp, L. and Sakamoto, H. (2021). 'Sober optimism and the formation of international environmental agreements', *Journal of Economic Theory*, vol. 197, 105321.

Keohane, R.O. and Oppenheimer, M. (2016). 'Paris: Beyond the climate dead end through pledge and review', *Politics and Governance*, vol. 4(3), pp. 42–51.
Kolstad, C.D. and Toman, M. (2005). 'The economics of climate policy', *Handbook of Environmental Economics*, vol. 3, pp. 1562–93.
Kovác, E. and Schmidt, R.C. (2021). 'A simple dynamic climate cooperation model', *Journal of Public Economics*, vol. 194, 104329.
Krishna, V. and Serrano, R. (1996). 'Multilateral bargaining', *Review of Economic Studies*, vol. 63(1), pp. 61–80.
Leonard, D. and Van Long, N. (1992). *Optimal Control Theory and Static Optimization in Economics*, Cambridge: Cambridge University Press.
Marchiori, C., Dietz, S. and Tavoni, A. (2017). 'Domestic politics and the formation of international environmental agreements', *Journal of Environmental Economics and Management*, vol. 81, pp. 115–31.
Martimort, D. and Sand-Zantman, W. (2016). 'A mechanism design approach to climate-change agreements', *Journal of the European Economic Association*, vol. 14(3), pp. 669–718.
Maskin, E. and Tirole, J. (2001). 'Markov perfect equilibrium: I. Observable actions', *Journal of Economic Theory*, vol. 100(2), pp. 191–219.
Nash, J. (1950). 'The bargaining problem', *Econometrica*, vol. 18, pp. 155–62.
Nash, J. (1953). 'Two-person cooperative games', *Econometrica*, vol. 21(1), pp. 128–40.
Nordhaus, W.D. (2015). 'Climate clubs: Overcoming free-riding in international climate policy', *American Economic Review*, vol. 105(4), pp. 1339–70.
OECD. (2018). 'Common time frames: Summary of discussions at the March 2018 climate change expert group global forum', Note prepared by the OECD/IEA Climate Change Expert Group.
Rubinstein, A. (1982). 'Perfect equilibrium in a bargaining model', *Econometrica*, vol. 50(1), pp. 97–109.
Schmalensee, R. (1998). 'Greenhouse policy architectures and institutions', in (W.D. Nordhaus, ed.), *Economics and Policy Issues in Climate Change*, pp. 137–58, Washington, DC: Resources for the Future Press.
Stern, N. (2006). *The Economics of Climate Change: The Stern Review*, Cambridge: Cambridge University Press.
Sydsaeter, K. and Hammond, P.J. (1995). *Mathematics for Economic Analysis*, Englewood Cliffs, NJ: Prentice Hall.
The New York Times. (November 28, 2015). https://www.nytimes.com/2015/11/29/opinion/sunday/what-the-paris-climate-meeting-must-do.html?searchResultPosition=1.
Tubiana, L. and Guérin, E. (2020). 'The Paris Agreement on climate change: What legacy?', in (C. Henry, J. Rockström and N. Stern, eds.), *Standing up for a Sustainable World*, pp. 103–15, Cheltenham: Edward Elgar Publishing.
Tirole, J. (2017). *Economics for the Common Good*, Princeton, NJ: Princeton University Press.
Victor, D. (2015). 'Why Paris worked: A different approach to climate diplomacy', *Yale Environment 360*, 15 December.