Pledge-and-Review Bargaining

27.8.2018. To request the most recent version, please email:

Bård Harstad

bard.harstad@econ.uio.no

Abstract

This paper analyzes a novel bargaining game, inspired by real-world negotiations. Before the Paris Agreement on climate change, each party submitted a nationally determined contribution, quantifying its contribution to a public good. The set of pledges was agreed to by everyone. The procedure will be repeated periodically, as newly developed technology makes earlier pledges obsolete.

I first show that, if the set of acceptable pledges is uncertain, each equilibrium pledge is approximated by an asymmetric Nash Bargaining Solution. The weights placed on others’ payoffs reflect the underlying uncertainty, but they vary from pledge to pledge, so the set of equilibrium pledges is inefficient.

This bargaining outcome is embedded in a dynamic contribution game, with endogenous participation, to investigate when the procedure is desirable. The results can rationalize the key differences between the climate agreements signed in Kyoto (1997) and Paris (2015) and the development from the former to the latter.

Key words: Dynamic games, bargaining games, the Nash Program, climate change, the Paris Agreement, the Kyoto Protocol
1 Introduction

*The pledge-and-review strategy is completely inadequate.*

Christian Gollier and Jean Tirole
The Economist (guest blog)
June 1st, 2015

Pledge-and-review bargaining refers to the structure of the negotiation process adopted in Paris, December, 2015. Before the countries were expected to sign the climate agreement, each party was asked to submit an "intended nationally determined contribution" (INDC). For most developed countries, the INDC specified unconditional 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

<table>
<thead>
<tr>
<th>Party:</th>
<th>Australia</th>
<th>Canada</th>
<th>EU</th>
<th>New Zealand</th>
<th>Norway</th>
<th>Russia</th>
<th>Switzerland</th>
<th>USA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pledge:</td>
<td>26-28%</td>
<td>30%</td>
<td>40%</td>
<td>30%</td>
<td>40%</td>
<td>25-30%</td>
<td>50%</td>
<td>26-28%</td>
</tr>
</tbody>
</table>

*Table 1. The pledges specify emission cuts relative to nationally chosen baselines.*

Every five years the parties shall review and make new pledges for another five-year period (Paris Agreement Art. 4).

This bargaining procedure is remarkably different from the one used under the Kyoto Protocol of 1997. There, a "top-down" approach was used to ensure the parties made legally binding commitments to cut emissions by (on average) five percent relative to the 1990-levels.² By comparison, pledge-and-review has been referred to as a "bottom-up" approach since countries themselves determine how much to cut nationally, without making these cuts conditional on other countries' emission cuts.³ No wonder, then, that economic theorists question the effectiveness of the pledge-and-review bargaining game.

This paper presents and analyzes a novel bargaining game, motivated by the structure of the Paris climate negotiations. I provide a characterization of the bargaining outcome and relate it to the Asymmetric Nash Bargaining Solution. I thereafter embed the bargaining solution into a dynamic game where parties can both contribute to a public good as well as invest in their future capacity to contribute

¹The baseline year is 1990 for the EU, Russia, and Switzerland, while it is 2005 for Australia, Canada, New Zealand, and USA. Several developing countries state pledges in terms of emission per GDP and some of these are conditional on receiving transfers. The official list is at http://www4.unfccc.int/ndcregistry but for an overview see http://cait.wri.org/indc/#/.

²For example, Bodansky and Rajamani (2017:11) write: "In essence, the Kyoto Protocol was the product of mutual concensions... The USA accepted a much stronger target (minus 7% from 1990 levels) than it had wanted..."

³As Victor (2017) writes: "Now, instead of setting commitments through centralized bargaining, the Paris approach sets countries free to make their own commitments." Paragraph 22 of the Paris Agreement "Invites Parties to communicate their first nationally determined contribution no later than when the Party submits its respective instrument of ratification, accession, or approval of the Paris Agreement. If a Party has communicated an intended nationally determined contribution prior to joining the Agreement, that Party shall be considered to have satisfied this provision."
(e.g., renewable energy sources), and I investigate how the bargaining procedure influences participation, compliance, and payoffs.

Remarkably, the difference in bargaining procedure (fact 1) can rationalize four other stylized facts on how the Paris Agreement differs from the Kyoto Protocol.\(^4\) (2) While relatively few (37) countries faced binding emission cuts under Kyoto, the Paris agreement has been signed by nearly every country in the world. (3) While the Kyoto Protocol was endogenously chosen in the 1990s, the participants preferred the pledge-and-review procedure in the 2010s. (4) While the commitments under the Kyoto Protocol was "legally binding," the INDCs are not. At the same time, and despite all these differences, (5) the commitment period length was five years for both treaties. I will now explain all this in detail.

(1) The novel feature of pledge-and-review (P&R) bargaining, the way I formalize it, is that each party is permitted to propose one’s own individual contribution only, rather than a vector of contributions for all the parties. I assume that all parties propose their pledges at the same time and, if at least one party rejects the set of pledges on the table, the process starts again after some delay.\(^5\) With complete information, the unique and trivial equilibrium of the game coincides with the non-cooperative (or "business-as-usual") outcome, where every party simply makes a pledge that maximizes one’s own utility. However, with sufficiently noisy shocks on the other parties’ willingness to decline and delay the agreement, I show that each party’s equilibrium contribution level coincides with the quantity that maximizes an asymmetric Nash product, where the weights on other parties’ payoffs reflect the extent of uncertainty as well as how the shocks are correlated. The weights on other parties’ payoffs are less than 1/2 for single-peaked and symmetric shock distributions, and they are (close to) zero when the variance of the shocks is small.

Note that this bargaining game is quite general and it might be useful for several other applications, beyond climate agreements. For example, the game may well capture a situation in which multiple business partners must negotiate a package, and where each partner is recognized as an expert and as the proposer for only one single dimension of the package: For example, one partner describes the product quality, another offers a strategy for advertisements, while a third manages a set of retailers, etc. In such expert meetings, it might be unrealistic to assume that a single partner is capable of proposing and describing all the details of interest, as is normally assumed in bargaining theory.

Although the bargaining game arguably has alternative applications, the assumptions are motivated by the pledge-and-review procedure, and thus the theory should be confronted with facts (2)-(5), described above. Most of the paper is therefore investigating the effects of the small weights associated with pledge and review. For this investigation, I present a dynamic game in which parties over time contribute to

\(^4\) These are the standard differences emphasized in the literature (see Bodansky and Rajamani 2017:11, or Pickering et al. 2017, for example).

\(^5\) If \(n\) sovereign countries contribute in equilibrium, the set of contributions must have been unanimously accepted by these \(n\) countries. Of course, it is possible for a subset of countries to sign treaties: Section 4 endogenizes \(n\) and Section 5 discusses minimum participation thresholds.
a public good (by making emission cuts) well as invest in their future capacities to contribute (e.g.,
they invest in green or renewable energy technology). The pledges quantify emission cuts and these are
revised and renegotiated periodically. Naturally, the small weights on others' payoffs, associated with
P&R, implies that emission cuts are not very ambitious, and thus investments in new technology, as well
as welfare, are lower than in the Nash Bargaining Solution. This (preliminary) negative result rationalizes
the critique of the pledge-and-review procedure.

(2) The negative result is reversed, however, when the decision to participate in the bargaining game
is endogenized. Since not much is expected from the participating countries (when the weights on others'
payoffs are small), it is not that costly for a party to participate, and this explains why the equilibrium
coalition size is larger with pledge-and-review bargaining. The larger coalition size implies that the sum
of contributions is larger, the aggregate investments larger, and so is welfare, I show.

(3) It is interesting to compare the outcome under P&R to the outcome following the NBS, often used
to describe the outcome of the Kyoto Protocol (cf. the literature review below). According to the results
described so far, the "shallow but broad" pledge-and-review game is superior because a larger number of
potential parties volunteer to participate. However, there is a limited number (π) of potential members
and if this upper boundary is binding, then the comparison is less clear. Further, when the parties are
heterogeneous, in that a number (n) of them will participate regardless of the game, then the "deep and
narrow" coalition under the Nash Bargaining Solution can be more attractive. With these constraints,
pledge-and-review is preferred if and only if n is large while π is small, I show.

This result is in line 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 twenty years, some of these have become emerging economies that potentially has an important
role to play. This implies that the number of relevant potential parties, π, has increased. During the same
period, seven of the countries that initially signed the Kyoto Protocol declared that they did not intend to
make commitments in the second commitment period of Kyoto: Belarus, Ukraine, Japan, New Zealand,
Russia, Canada, and USA. This can be interpreted as a smaller n. Either (or both) of these developments
makes pledge-and-review more attractive. Thus, the theory can rationalize why the top-down approach
characterizing the Kyoto Protocol was chosen in the 1990s, while pledge-and-review was preferred in the
2010s.

(4) If the parties cannot commit to future actions, the pledges must be self-enforcing. As in the
repeated games literature, one may require a party to be willing to comply if the alternative is that
cooperation (eventually) breaks down. The P&R bargaining outcome is more likely to be self-enforcing
than the NBS, I show. The main reason for this result is that when the negotiated pledges are less
ambitious, the temptation to defect is small.

If the bargaining outcome is characterized by the NBS, in contrast, the parties might find it necessary
to motivate compliance by raising the political cost of defection. In practice, the political cost can be
raised by requiring the emission cuts to be "legally binding" or enforced by other punitive measures—and the Kyoto Protocol is indeed referring to both these measures.

(5) 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 new capacity, 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 it is independent of the number of participants, I show. Thus, if a five-year commitment period was optimal under the Kyoto Protocol, it is indeed optimal also for the Paris Agreement, according to this theory.

**Literature (to be updated).** The pledge-and-review bargaining game has not been analyzed in the theoretical literature, as far as I know. By showing that this bargaining game implements an asymmetric (or "generalized") Nash Bargaining Solution (NBS) for each party’s contribution, I contribute to the 'Nash Program', aimed at finding noncooperative games implementing cooperative solution concepts. The Nash demand game, first described by Nash (1953), intended to implement the Nash Bargaining Solution, axiomatized by Nash (1950). There is a large subsequent literature investigating the extent to which the Nash demand game implements the Nash Bargaining Solution (Binmore, 1992; Abreu and Gul, 2000; Kambe, 2000), and some contributions also allow for uncertainty, as I do here (Binmore, 1987; Carlsson, 1991; Andersson et al., 2017).

The alternating offer bargaining game by Rubinstein (1982) also implements the Nash Bargaining Solution, as shown by Binmore et al. (1986). Although there can be multiple equilibria with more than two players (Sutton, 1986; Osborne and Rubinstein, 1990), reasonable consistency conditions support the NBS in multi-lateral negotiations (Krishna and Serrano, 1996; see also Jun, 1987, and Chae and Yang, 1988 and 1994). For these reasons, the NBS is often, and reasonably, assumed to characterize the outcome of top-down negotiations, such as the Kyoto Protocol (Beccherle and Tirole, 2011; Battaglini and Harstad, 2015; Harstad, 2012, 2016). In contrast, the asymmetric Nash Bargaining Solution (ANBS; axiomatized by Harsanyi and Selten, 1972; Kalai, 1977; Roth, 1979) characterizes the outcome if there are asymmetric discount rates or recognition probabilities (Miyakawa, 2008; Britz et al. 2010; Laruelle and Valenciano, 2008), in addition to the discount rates (Kawamori, 2014).

The next section contributes to this literature by showing that also 'pledge-and-review' bargaining implements an asymmetric NBS for each party’s contribution. However, the weights are shown to vary from one party’s contribution level to another’s, so the set of contributions is not Pareto optimal. The weights will reflect differences in the discount rates (as in some of the papers already mentioned), but also the extent of uncertainty in shocks and the correlation in shocks across the parties.

---

6 Also the Nash bargaining solution with endogenous threats has been given noncooperative foundations in dynamic games (Abreu and Pearce, 2007 and 2015).

7 There are also papers showing how the NBS is implemented exactly in other ways, either by a specific game (Howard, 1992) or in a matching context (Cho and Matsui, 2013).
The coalition formation game is the standard game used to model collusion (d’Aspremont et al., 1983) or environmental coalitions (Hoel, 1992, Carraro and Siniscalco, 1993, Barrett, 1994). This literature predicts that the equilibrium coalition size is very small because of the free-riding incentives. Kolstad and Toman (2005) and Nordhaus (2015) thus refer to it as a paradox that actual environmental coalitions can often be quite large. There is, however, a well-known trade-off between treaties that are narrow-but-deep (as in the above-mentioned papers) vs. broad-but-shallow (Finus and Maus, 2008; IPCC, 2014). My contribution to this literature is to show why some bargaining procedures (such as pledge-and-review) lead to shallow (and thus broad) coalitions, and to show when this procedure is chosen in equilibrium. I also contribute to the literature on self-enforcing agreements (see, for example, Barrett, 1994, Dutta and Radner, 2004, or Harstad et al., 2018) by showing when and why certain procedures, such as pledge-and-review, are more likely than others to be self-enforcing.

Outline. The next section formalizes pledge-and-review bargaining and characterizes the outcome in Theorem 1 and 2. Section 3 embeds the bargaining outcome in a dynamic climate/contribution game in which parties over time cut emissions as well as invest in their capacities to cut emissions in the future: P&R leads to lower contributions, investments, and welfare, but this negative finding is reversed when participation is endogenous, in Section 4. Section 5 argues that the theory can explain why the signatories have switched from preferring the NBS (in the 1990s) to P&R (in the 2010s), while Section 6 shows that the P&R pledges are more likely to be self-enforcing, without being 'legally binding.' Section 7 shows that the optimal commitment period is the same for the two bargaining procedures. Ten extensions are briefly discussed in Section 8 before Section 9 concludes. The Appendix contains all proofs.

2 A Model of Pledge-and-Review Bargaining

This section describes a novel bargaining game, not yet analyzed in the literature, and characterizes its outcome. The section may be read independently from the other sections, as the model here may have alternative applications than climate negotiations. As mentioned, the bargaining game might be appropriate when a number of business partners are negotiating a multi-dimensional deal, and each partner has expertise on and is making the proposal on one single dimension of the package (such as quality, price, delivery time, etc). The main new feature of the game is that each party is recognized as being responsible for proposing only one dimension of the agreement, even though payoffs depend on the entire vector.

---

8These coalition formation games are rather simple, although the effect of far-sightedness has also been discussed (Ray and Vohra, 2001).
2.1 The Bargaining Game

There are $n$ parties, each endowed with a payoff function $U_i : \mathbb{R}^n \rightarrow \mathbb{R}$, $i \in \{1,...n\}$. The bargaining game starts when each party $i$ simultaneously proposes its own dimension, or contribution, $x_i \in \mathbb{R}$, before observing the vector of proposed contributions, $\mathbf{x} = (x_1,...,x_n)$. Thereafter, each party must decide whether to accept (or ratify) the proposed agreement, $\mathbf{x}$. If one or more party declines $\mathbf{x}$, the game restarts after some finite delay, $\Delta \in (0, \infty)$. If everyone accepts, every party $i$ receives the payoff $U_i(\mathbf{x})$ and the game ends.\footnote{Two remarks are in order: (i) It is not assumed that the $x_i$'s are enforced; only that they influence (continuation) payoffs. (ii) Unanimity is required among the parties that (are expected to) contribute positively, but the $n$ parties can be a subset of all players (see Sections 4 and 5).}

I assume $U_i$ to be concave and continuously differentiable, and both $U_i$ and $x_i$ are measured relative to the default outcome (which is therefore normalized to zero). Furthermore, I will start by making the additional assumptions $\partial U_i(\cdot)/\partial x_i < (>) 0$ for $x_i > (<) 0$, and $\partial U_j(\cdot)/\partial x_i > 0$, $j \neq i$, so that the $x_i$'s can be interpreted as contributions to a public good. However, the Appendix proves Theorem 2 and a generalization of Theorem 1 without these additional assumptions: see the below 'Remarks on generality'.

Party $i$'s discount factor between time $t$ and $t+\Delta$ is $\delta^\Delta_{i,t} \leq 1$, but it will be more convenient to refer to the "discount rate" $\rho_{i,t} \equiv \left(1 - \delta^\Delta_{i,t}\right)/\Delta$.\footnote{If the real discount rate is $\tilde{\rho}_{i,t}$, the discount factor is $e^{-\tilde{\rho}_{i,t}\Delta} = \delta^\Delta_{i,t}$, so $\rho_{i,t} \equiv \left(1 - e^{-\tilde{\rho}_{i,t}\Delta}\right)/\Delta$ approaches $\tilde{\rho}_{i,t}$ when $\Delta \rightarrow 0$. I refer to $\rho_{i,t}$ as the discount rate even though the identity holds only in the limit.} Thus, $i$ receives $(1 - \rho_{i,t}\Delta)U_i(\mathbf{x}^*)$ by declining an offer if $\mathbf{x}^*$ can be expected next period. Given $\mathbf{x}^*$, $i$ prefers to accept $\mathbf{x}$ now if:

$$U_i(\mathbf{x}) \geq (1 - \rho_{i,t}\Delta)U_i(\mathbf{x}^*). \quad (1)$$

If information were perfect, and if each $\rho_{i,t} = \rho_i > 0$, it is easy to see that $\mathbf{x}^*$ could be a stationary SPE only if $x_i^* = \arg \max_{x_i} U_i(x_i, \mathbf{x}_{-i}^*)$, if $U_j(\mathbf{x}^*) > 0 \forall j$. For any other equilibrium candidate, $i$ could always suggest an $x_i$ slightly different from $x_i^*$ without violating (1). Therefore, with the additional assumptions above, the "trivial equilibrium" $\mathbf{x}^* = \mathbf{0}$ would be unique. This observation confirms the pessimism associated with pledge and review, as described in the Introduction.

In reality, party $i$ is unlikely to know precisely the condition under which an offer will be accepted. One way of modelling this uncertainty is to assume that the exact discount rates for the next period are not known (to anyone) at the time at which the offers are made. After all, a decision maker’s impatience when it comes to accepting an agreement may depend on a range of temporary domestic policy or economy issues. To capture this, write $\rho_{i,t} = \theta_{i,t}\rho_i$, where $\rho_i$ is $i$'s expected discount rate while $\theta_{i,t}$ is a shock with mean 1. The shocks are jointly distributed with pdf $f(\theta_{1,t},...,\theta_{n,t})$ on support $\prod_i [0, \bar{\theta}_i]$, i.i.d. at each time $t$, and the marginal distribution of $\theta_{i,t}$ is $f_i(\theta_{i,t}) \equiv \int_{\Theta_{-i}} f(\theta_{1,t},...,\theta_{n,t})$, where $\Theta_{-i} \equiv \prod_{j \neq i} [0, \bar{\theta}_j]$. The $\theta_{i,t}$'s are realized and observed by everyone after the offers but before acceptance decisions are...
made.\textsuperscript{11} Since it follows that the game is stationary, I will henceforth restrict attention to stationary subgame-perfect equilibria (SPEs).

2.2 The Bargaining Solution

After learning $\theta_{i,t}$, $i$ accepts $x$ if and only if:

$$U_i(x) \geq (1 - \theta_{i,t} \rho_i \Delta) U_i(x^*) \Rightarrow \theta_{i,t} \geq \frac{U_i(x^*) - U_i(x)}{\rho_i \Delta U_i(x^*)}. \quad (2)$$

When $\theta_{i,t}$ is drawn from a continuous distribution, the probability that $i$ accepts will be continuous in $x_i$. As the following result will show, this continuity can motivate larger contributions: $x^*$ can be supported as a "nontrivial" stationary SPE if the marginal benefit for $i$ by slightly reducing $x_i$ is outweighed by the risk that at least one party might be sufficiently patient to decline the offer and wait for $x^*$.

**Theorem 1:** If $x^*$ is a nontrivial stationary SPE in which $U_i(x^*) > 0 \forall i$, then, for every $i \in N$:

$$x^*_i \leq \arg \max_{x_i} \prod_{j \in N} (U_j(x_i, x^*_{-i}))^{w_{i,j}}, \text{ where } \frac{w_{i,j}}{w_i} = \frac{\rho_i}{\rho_j} f_j(0) E(\theta_{i,t} | \theta_{j,t} = 0), \forall j \neq i. \quad (3)$$

Before explaining and discussing the result, note that the above reasoning does not limit how small the equilibrium $x^*_i$'s can be, as there is no point for $i$ to contribute more than $x^*_i$, whatever the equilibrium $x^*$ is. (The stationary equilibrium $x^*$ will always be accepted, as is evident from (2) given that $\theta_{i,t} \geq 0$). There can thus be multiple equilibria. To obtain sharper results, it is common to require the equilibrium to be robust to small trembles. If there is a risk that even $x^*$ will be declined, then $i$ may prefer to reduce the risk until (3) binds. This logic holds if we impose the following version of trembling-hand perfection.

**Definition of Small Trembles.** When the intended offers are given by $x$, $x + \epsilon_k^t$ is realized, where $\epsilon_k^t$ is a vector of $n$ shocks, each i.i.d. over time with mean zero and $E(\epsilon_k^t)^2 \to 0$ as $k \to 0$.

**Theorem 2:** If $x^*(k)$ is a nontrivial stationary SPE with Small Trembles, then (3) holds with equality for $x^*_i = \lim_{k \to 0} x^*_i(k)$, for every $i \in N$.

As a comparison, note that if $x$ were given by the Asymmetric Nash Bargaining Solution (ANBS), then the outcome could be described as:

$$x^*_i = \arg \max_{x_i} \prod_{j \in N} (U_j(x_i, x^*_{-i}))^{w_{i,j}}. \quad (4)$$

\textsuperscript{11}This is not unreasonable: (i) Technically, instead of letting $\Delta > 0$ be the delay between rejections and new offers, $\Delta$ can be the delay between offers and acceptance decisions, if we assume that new offers can be made as soon as earlier offers are rejected. (ii) Since there is (then) a lag between offers and acceptance decisions, it is natural that policy makers in the meanwhile learn about how urgent it is for them to conclude the negotiations, or about the attention they instead have to give to other policy and economic issues.
In the ANBS, each $x_i^A$ maximizes the Asymmetric Nash product, and thus a weighted sum of utilities, where the weights $w_j$’s are exogenously given. In this case, the set of $x_j^A$’s will be Pareto optimal.

Also when (3) binds, following pledge-and-review bargaining, the equilibrium $x_i^*$ maximizes an asymmetric Nash product, but the weights vary with $i$ and thus the set of $x_i^*$’s is not Pareto optimal. In particular, if $w_j^i/w_i^i < 1$ for every $(i,j), j \neq i$, then it is possible to make every party better off by increasing all contributions relative to $x^*$.

Theorem 1 also endogenizes the weights and shows how they depend on three things. First, the weights on $j$’s utility is larger if $j$ is expected to be patient relative to $i$. This is natural (and in line with other bargaining papers, as discussed in the Introduction): When $j$ is patient, $j$ is more tempted to reject an offer that is worse than what one can expect in the next period, and thus $i$ finds it too risky to reduce $x_i$, especially when $i$ is quite impatient and dislikes delay.

Second, the weight on $j$’s payoff is larger when there is a lot of uncertainty regarding $j$’s shock. Of importance is especially the (marginal) likelihood that $j$’s discount rate is close to 0, so that even a small reduction from $x_i^*$ involves some risk that $j$ will decline.

Third, if the shocks are correlated, then the weight on $j$’s payoff is less for a small $\mathbb{E}(\theta_i | \theta_j = 0)$, which measures $i$’s expected shock (on the discount rate) given that $j$’s shock is small. Intuitively, if $i$ can be expected to have a small discount rate exactly when $j$ has, then it matters less that $j$ declines an offer in this circumstance. When the delay matters less, $i$ does not find it necessary to offer a lot. A party $i$ will therefore pay more attention to the payoffs of those other parties who face shocks that are less correlated with $i$’s shock. In sum, each party pledges to contribute an amount that puts some weight on the utility of other parties, but only to the extent that one is uncertain about the other’s willingness to accept.

### 2.3 Implications and Generalizations

The theorems have several important consequences.

**Corollary 1:** Suppose all parties have the same preferences and shock distributions, $f$.

(i) In equilibrium, the $x_i^*$’s are given by:

\[
x_i^* = \arg \max_{x_i} U_i(x_i, x_{-i}^*) + w \sum_{j \neq i} U_j(x_i, x_{-i}^*), \quad \text{where}
\]

\[
w = f(0) \mathbb{E}(\theta_i | \theta_j = 0) \forall i, j.
\]

(ii) The weight is $w \leq 1/2 \forall i, j$, if $f(\cdot)$ is single-peaked and symmetric and shocks are not negatively correlated.\(^{12}\)

\(^{12}\)To see this, note that if $f(0) > 1/2$, then, when $f(\cdot)$ is single-peaked and symmetric around the mean of one, \[\int_0^1 f(\theta_j) d\theta_j > 1,\] violating the definition of a pdf. If the shocks are i.i.d., then $\mathbb{E}(\theta_i | \theta_j = 0) = 1$. If shocks are positively correlated, then $\mathbb{E}(\theta_i | \theta_j = 0) \leq 1$. \]
Figure 1: There are multiple equilibrium contribution levels in Example E, but they are all smaller than both the efficient levels ($x^S$) and the unique nontrivial equilibrium when there are Small Trembles ($x^*$).

Combining the two parts, the corollary suggests that the weight on other parties’ payoffs is less than 1/2 of the weight on $i$’s payoff when $x_i$ is proposed, if preferences are similar. If uncertainty vanishes, such that the pdf $f_j(.)$ concentrates around its mean, then $f_j(0) \to 0$, so $w \to 0$, and $x_i^*$ must approach the level in the trivial equilibrium as when there is no bargaining.

**Example E.** As an illustration, consider the situation in which $i$ benefits (linearly) from the other’s contributions, while individual contributions has a quadratic cost:

$$U_i(x_i, x^*_{-i}) = \alpha \sum_{j \neq i} x_j - \beta x_i^2/2 + \gamma,$$  \hspace{1cm} (E)

for some positive constants $\alpha$, $\beta$, and $\gamma$. Corollary 1 implies:

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

The colored area in Figure 1 illustrates the set of equilibria (without Small Trembles) when $n=2$ and with symmetric $w_j^i/w_i^i = w^\forall i, j$, for some $w < 1$. (The pair of dashed curved lines corresponds to $w = 1$, and the symmetric Nash Bargaining Solution is illustrated by $x_1^S$ and $x_2^S$.) In this example, it is easy to check that all equilibria satisfying (3) with strict inequalities are Pareto dominated by the equilibrium where the inequalities bind (i.e., $x_1^*$ and $x_2^*$) if just $w < \sqrt{3} - 1 \approx 0.73$. Thus, focusing on equilibria that are not Pareto dominated can in some cases replace the assumption on the trembles.\(^{13}\)

**Remarks on generality.** (i) The inequality in (3) will bind for nontrivial equilibria also without the

\(^{13}\)I thank Asher Wolinsky for making this observation.
Small Trembles, if we, as an alternative, introduce trembles on the supports of the discount rates.\footnote{In Theorem 2, the assumption on Small Trembles can be replaced by introducing trembles on the support for the discount rate, i.e., if the support of $\theta_{j,t}$ were $[\underline{\theta}_j^k, \bar{\theta}_j]$ (instead of $[0, \bar{\theta}_j]$), where $\underline{\theta}_j^k < 0$ and $\underline{\theta}_j^k \uparrow 0$ as $k \to 0$. The interpretation of a negative discount rate may be that, in some circumstances, a party prefers to delay signing agreements due to other urgent economic/policy issues that requires the decision makers' attention. It is required that the lower boundaries, the $\underline{\theta}_j^k$'s, approach zero in the limit (as $k \to 0$), since otherwise there will be delay on the equilibrium path.}

(ii) Note that the theorems do not require $\Delta$ to be small. (iii) Although I assumed above that $\partial U_i(\cdot) / \partial x_i < 0$ and $\partial U_j(\cdot) / \partial x_i > 0$, $j \neq i$, these assumptions are not needed for Theorem 2, and a generalization of Theorem 1 is proven in the Appendix without these additional assumptions. (iv) If the trivial equilibrium gives strictly positive payoffs to at least one party, then it ceases to exist when we assume the Small Trembles and, therefore, the word "nontrivial" in Theorem 2, is, in this case, redundant.\footnote{This explains why the qualifier in Theorem 1, "in which $U_i(x^*) > 0$", is absent from Theorem 2.}

**Remark on sufficiency.** Condition (3) is necessary for $x^*$ to be an equilibrium, but it may not be sufficient. Whether the second-order condition for an optimal deviation for $i$ holds globally depends on the pdf's $f_j$, $j \neq i$. If $n = 2$, a sufficient condition for the second-order condition to hold is that $f_j$ is weakly increasing, as when $\theta_{j,t}$ is uniformly distributed, for example.\footnote{For this case, it is easy to see from the first-order condition in the proof, equation (26), that the second-order condition is satisfied.}

2.4 Remarks on Nash’s Demand Game and Bargaining Solution

The P&R bargaining outcome is in stark contrast to the symmetric Nash Bargaining Solution, predicting that the $x_i$’s would follow from (4) with $w = 1$. The NBS is frequently used to describe multilateral bargaining outcomes, such as the Kyoto Protocol, because it results from standard bargaining games, as explained in the Introduction: (1) The NBS is the stationary noncooperative equilibrium in an alternating offer bargaining game when the proposer specifies a complete outcome.\footnote{The literature following Binmore et al. (1986) shows that the stationary SPE of an alternating offer bargaining game, where each party is equally likely to be recognized as the proposer at the beginning of each period, approaches the NBS when $\Delta \to 0$ and discount rates are equal. With different discount rates, the equilibrium approaches the ANBS where the difference in discount rates, $\rho_j / \rho_i$, characterizes $w_i / w_j$, i.e., the bargaining weight on party $i$ relative to party $j$. In this case, a party takes into account the payoff of another part because the other party may be the proposer in the next round. This incentive (and the bargaining outcome) would be unchanged if $i$ also had to take into a fraction of $j$’s payoff because of the uncertainty introduced above. The uncertainty in the discount rates, introduced above, will not change these results.}

(2) The NBS is also the noncooperative outcome of the Nash Demand Game (NDG). Interestingly, result (2) follows as a corollary to Theorem 2 because the NDG is a special case of P&R.

In the NDG, each player is demanding an (ex post) utility level or, equivalently, a variable ($x_i$) that dictates $i$’s ex post utility, $v_i(x_i)$. The vector of demands is feasible with probability $P(x)$, so $i$’s expected payoff is:\footnote{Binmore (1987) offers a more detailed description and analysis of this game.}

$$U_i(x_i, x_{-i}) = v_i(x_i) P(x). \tag{5}$$

When (5) is substituted into a binding (3), the online appendix proves:
Corollary 2: Suppose every payoff is given by (5). With PR bargaining, Theorem 2 implies:

\[ x^* = \arg \max_x \prod_{i \in N} v_i(x_i)^{q_i} P(x)^{\bar{c}} \]

\[ = \arg \max_x \prod_{i \in N} v_i(x_i)^{q_i} \text{ s.t. } P(x) = P(x^*) \text{, where} \]

\[ q_i = \frac{w_i^j / \sum_j w_j^i}{\sum_k \left( \frac{w_k^j / \sum_j w_j^k}{\sum_j w_j^k} \right)} \quad \text{and} \quad \bar{c} = \frac{1}{\sum_k \left( \frac{w_k^j / \sum_j w_j^k}{\sum_j w_j^k} \right)}. \]

Note that \( x^* \) coincides with the NBS in a special case, namely if (a) the \( w_j^i \)'s are the same for all and (b) uncertainty vanishes, in the sense that \( P(x) \) is close to 0 or 1 for most \( x \)'s. With (b), it is intuitive that \( x^* \) must be close to an \( x \) that ensures \( P(x) \approx 1 \), and the constraint \( P(x) = P(x^*) \) in (7) is simply requiring \( x \) to be feasible. Assumption (a) holds if, for example, \( f_j(0) = 0 \quad \forall j \), as in the traditional literature. In the Nash Demand Game, there is neither voting nor additional rounds, so party \( i \) does not need to fear that the others will reject the proposal, just as in the P&R game when \( f_j(0) = 0 \quad \forall j \). The mapping from the NDG to the NBS is in several respects generalized by Corollary 2 and Theorem 2.\(^{19}\)

If uncertainty is not vanishing, so that \( P(x) \) is a (more) continuous function, then (6) shows that each party places larger weights on the (collective) risk if the \( w_j^i \)'s are large; this increases the equilibrium probability that the set of demands is compatible. Similarly, the function \( U_i(x) \) is a continuous in the \( x_j \)'s also if the parties negotiate over individual contributions, as in Example E, rather than utility levels. Then, the outcome is inefficient with P&R bargaining, especially when the \( w_j^i \)'s are small.

3 A Dynamic Contribution Game

To better understand the implications of pledge-and-review, this section embeds the pledge-and-review bargaining solution into a tractable dynamic game with externalities. The model describes a dynamic contribution game (to a public good) in which the parties over time can contribute as well as invest in their 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 analyze climate policies in a dynamic setting. Chapter 16 of the Stern Review (2007) pointed out that new technology would be crucial to mitigate climate change. At the same time, §114 of the 2010 Cancun Agreement states that “technology needs must be nationally determined, based on national circumstance and priorities.” Thus, to be consistent with traditional climate change negotiations, emission cuts are negotiable (and contractible), while technology investments are not (Section 8 relaxes this assumption).

\(^{19}\)(i) Theorem 2 allows for a finite delay if \( x^* \) is not approved unanimously, before the game can be plaid again, as well as shocks on the discount rates. Despite these differences, the outcome continues to be the NBS if the parties are similar, regardless of the level of the common \( w \). The intuition for why \( w \) is irrelevant is that, given the sharp threshold characterized by \( P(x) \) when uncertainty vanishes, \( i \)'s preferred \( x_i \) coincides with the efficient level, given the other \( x_j \)'s.

(ii) When the parties are heterogeneous, the outcome (7) is given by an ANBS. The bargaining power index \( (q_i) \) is larger for those parties who are likely to be patient or face less uncertainty regarding the opponents’ discount rates.
3.1 A Model of Climate Policy

For the dynamic model to be tractable, I restrict attention to linear-quadratic per-period utility functions. At each time \( t \), the utility for party \( i \) is the sum of three parts. First, if each party \( i \) contributes \( x_{i,t} \), the benefit from the sum of contributions is \( a \sum_j x_{j,t} \). This linearity assumption is made for simplicity, but it may also reflect the fact that the marginal benefit from \( i \)'s contribution at time \( t \) is unlikely to change dramatically over (short periods of) time. For example, Golosov et al. (2014) estimate that the climate damage function is approximately 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, since \( a \) 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 party \( i \) emits \( g_{i,t} \) and the pollution stock is \( G_t = G_{t-1} + \sum_j g_{j,t} \), where \( \sigma \in [0, 1] \) measures the fraction of the past stock that survives to the next period. If parameter \( C > 0 \) measures each party’s per-period marginal environmental harm from the stock \( G_t \), then the present-discounted harm of another unit of emission is \( C/(1 - \sigma \delta) \) for each party, which we can define as \( a \). Consequently, \( a \equiv C/(1 - \sigma \delta) \) also measures the present-discounted benefit from a marginal cut in the current emission level.

The second term in the utility function specifies the cost of contributing to the public good. To fix ideas, suppose a country can consume energy from both fossil fuels \( (g_{i,t}) \) as well as renewables \( (\gamma_{i,t}) \). If the total consumption of energy is less than \( i \)'s bliss point, \( \gamma_i \), then \( i \) experiences a disutility that is quadratic in the difference: \( \frac{b}{2} \left( \gamma_i - \left[ g_{i,t} + \gamma_{i,t} \right] \right)^2 \). This disutility can be written as \( \frac{b}{2} \left( \tilde{x}_{i,t} - \gamma_i \right)^2 \), when \( \tilde{x}_{i,t} \) represents a cut in emission relative to \( i \)'s bliss point (i.e., when \( \tilde{x}_{i,t} = \gamma_i - g_i \)).

Of course, also for other public good situations, it is natural that it is particularly costly for \( i \) to contribute a lot relative to \( i \)'s capacity level, measured by the stock \( \gamma_{i,t} \).

Each party can over time add to the capacity \( \gamma_{i,t} \) by investing \( \bar{y}_{i,t} \). The investment cost is assumed to be convex and quadratic and it constitutes the third term in the per-period utility function:

\[
\bar{u}_{i,t} = a \sum_j \tilde{x}_{j,t} - \frac{b}{2} \left( \tilde{x}_{i,t} - \gamma_i \right)^2 - \frac{c}{2} \bar{y}_{i,t}^2, \text{ where} \tag{8}
\]

\[
\gamma_{i,t+1} = \gamma_{i,t} + \bar{y}_{i,t}, \quad \text{and where} \ a, \ b, \text{ and } c \text{ are positive constants.} \]

The parties can have heterogeneous bliss points for consumption and initial technology levels \( (\gamma_{i,0}) \). For simplicity, the parties are assumed to be identical in other respects and they plan by applying the same (expected) discount factor, \( \delta \).\[20\]

---

\[20\] Thus, \( i \) seeks to maximize \( \sum_{t=0}^{\infty} \delta^t \bar{u}_{i,t} \) at time 0. Although a party’s impatience was allowed to be stochastic and uncertain during the bargaining process in Section 2 (in order to make the parties’ acceptance decisions uncertain), I henceforth assume the parties apply the same constant and deterministic (expected) discount factor when they decide on the long-lasting investment levels. This is natural, since the uncertainty in the willingness to accept bargaining offers could
**BAU:** As a benchmark, consider the noncooperative MPE without any treaty, referred to as the "business as usual" (BAU) equilibrium. At every time $t$, when $i$ takes as given $\bar{Y}_{i,t}$, the marginal benefit of emitting equals the marginal cost for party $i$:

$$b(\bar{x}_{i,t} - \bar{Y}_{i,t}) = a \Rightarrow \bar{x}^{BAU}_{i,t} = \bar{Y}_{i,t} + \frac{a}{b}.$$  

Consequently, the investment level does not influence $i$’s future consumption level $(\bar{x}_{i,t} - \bar{Y}_{i,t})$, but only $i$’s contribution levels in every future period. The investment level is thus:

$$\bar{y}^{BAU}_{i,t} = \frac{\delta}{1 - \delta} \frac{a}{c}.$$  

It is straightforward to derive party $i$’s continuation value in BAU, $U_{i,t}^{BAU}$, given the constant $\bar{y}^{BAU}_{i,t}$ and $\bar{y}^{BAU}_{i,t}$.\(^{21}\)

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

**Pledges:** Now, consider the possibility that the parties will contribute more than the BAU-levels. In particular, suppose $i$ agrees at $t = 1$ to contribute $x_i \geq 0$ units, beyond $i$’s BAU level, for each of the next $T$ periods.\(^{22}\) I will continue to restrict attention to Markov-perfect strategies, just as for the BAU scenario.\(^{23}\) However, the commitment $x_i$ is payoff-relevant and it might motivate $i$ to invest $y_{i,t}$ units in addition to the business-as-usual level. Total contributions and investments can then be written as:

$$\bar{x}_{i,t} = \bar{x}^{BAU}_{i,t} + x_i \text{ and } \bar{y}_{i,t} = \bar{y}^{BAU}_{i,t} + y_{i,t}. \quad (10)$$

The analysis will focus on the choices of $x_i$ and $y_{i,t}$, since these also will pin down $\bar{y}_{i,t}$ and $\bar{y}_{i,t}$, given the constant $\bar{y}^{BAU}_{i,t}$ and $\bar{y}^{BAU}_{i,t}$.

**Timing:** The timing of the game is the following. The parties negotiate the contributions (relative to BAU), $x_i$, pinning down the contribution levels for the next $T$ periods $\{1, ..., T\}$. Thereafter, in every period during the commitment period, each party decides on its investment level. After $T$ periods, the parties negotiate new contributions according to pledge-and-review, once again.\(^{24}\)

---

\(^{21}\) As proven in the Appendix:

$$U_{i,t}^{BAU} = \frac{a}{1 - \delta} \sum_j \bar{Y}_{i,t} + \frac{a^2}{1 - \delta} \left( n - \frac{1}{2} \right) \left( \frac{1}{b} + \frac{1}{c} \left[ \frac{\delta}{1 - \delta} \right]^2 \right).$$

\(^{22}\) Note that the additional contribution, $x_i$, is assumed to be constant over the commitment period. However, Section 8 explains why the results below are robust and continue to hold if, for example, the contributions are time-dependent.

\(^{23}\) There are a number of good reasons for why parts of the literature restrict attention to MPEs. That said, Section 6 permits history-contingent strategies to explore when the equilibrium agreement is self-enforcing.

\(^{24}\) The pledges in the Paris Agreement will be renewed every five years. According to Carbon Brief, “The idea is that this short time-frame would give countries the opportunity to regularly capture scientific and technological developments in their official targets, as well adjust for any overachievement.” [https://www.carbonbrief.org/explainer-the-ratchet-mechanism-within-the-paris-climate-deal](https://www.carbonbrief.org/explainer-the-ratchet-mechanism-within-the-paris-climate-deal)
After reformulating the problem, Section 3.3 derives the equilibrium investment levels, for each point in time, as a function of the \( x_i \)'s. This function implies that the continuation values can be summarized as a function of the \( x_i \)'s. Thus, the theorems can be applied to characterize the bargaining outcome for the \( x_i \)'s and the effects of pledge-and-review on equilibrium investment levels.

**Extensions:** I start by treating as exogenous parameters \( n, T, \) and \( w \), since these might be determined by forces outside of this model, but \( n \) is endogenized in Section 4, and Section 5 discusses the preference over \( w \) (i.e., when pledge-and-review bargaining is preferred to the Nash Bargaining Solution, for example). Section 6 analyzes compliance and the conditions under which the pledges are self-enforcing, while Section 7 discusses the optimal \( T \). Section 8 explains that the results also hold with alternative timings than the one illustrated in Figure 2 and also if the pledges are functions of time, relate to the investments levels or emission taxes, or if the investments are made by firms rather than governments.

### 3.2 The Optimal Control Problem

Party \( i \)'s continuation value can be written as a function of the \( x_i \)'s and the \( y_{i,t} \)'s, given (10) and the BAU constants. After the pledges have been agreed to, party \( i \)'s problem is to choose the investment levels over the next \( T \) periods.

**Lemma 1:** Given the actual pledges, \( x \), and the future equilibrium pledges, \( x^* \), party \( i \)'s continuation value is \( U^B_i + U_i(x) \), where:

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

\[
Y_{i,t+1} = Y_{i,t} + y_{i,t} , \text{ and } Y_{i,1} = 0.
\]

The lemma permits the current pledges \( (x) \) to be different from those that are expected in equilibrium in the subsequent commitment period (i.e., \( x^* \)).

Conveniently, the heterogenous bliss points and the initial technology levels drop out when utility is measured relative to BAU. As in the BAU case, is irrelevant that pollution is a stock that accumulates over time, as long as \( a \) measures the present-discounted value of abating one unit.
Rather than starting with the climate-change motivation in the previous subsection, the reader could start directly by considering the problem in Lemma 1. Here, as in several other public-good settings, party \( i \) benefits from the others’ contributions but faces a cost of contributing more than \( i \)'s capacity.

### 3.3 Equilibrium Investments

The exact solution for the investment and the technology levels is presented in the following lemma.

**Lemma 2:** For every \( i \in N \) and \( t \in \{1, ..., T\} \), the equilibrium investment \( y_{i,t} \) and stock \( Y_{i,t} \) are linear in \( x_{i} \):

\[
\begin{align*}
y_{i,t} &= x_{i} (C_{1}L_{1}^{t-1} [1 - L_{1}] - C_{2}L_{2}^{t-1} [L_{2} - 1]) \quad \text{and} \\
Y_{i,t} &= x_{i} (1 - C_{1}L_{1}^{t-1} - C_{2}L_{2}^{t-1}) ,
\end{align*}
\]

where

\[
\begin{align*}
L_{1} &= \frac{1}{2} \left( \frac{1}{\delta} + 1 + \frac{b}{c} \right) - \frac{1}{2} \sqrt{\left( \frac{1}{\delta} + 1 + \frac{b}{c} \right)^2 - \frac{4}{\delta}} \in (0, 1) \\
L_{2} &= \frac{1}{2} \left( \frac{1}{\delta} + 1 + \frac{b}{c} \right) + \frac{1}{2} \sqrt{\left( \frac{1}{\delta} + 1 + \frac{b}{c} \right)^2 - \frac{4}{\delta}} > 1, \\
C_{1} &= \frac{L_{2}^{T-1} (L_{2} - 1)}{L_{1}^{T-1} (1 - L_{1}) + L_{2}^{T-1} (L_{2} - 1)} \in (0, 1), \\
C_{2} &= \frac{L_{1}^{T-1} (1 - L_{1})}{L_{1}^{T-1} (1 - L_{1}) + L_{2}^{T-1} (L_{2} - 1)} = 1 - C_{1} \in (0, 1).
\end{align*}
\]

Naturally, if \( i \) is committed to contribute a lot, in that \( x_{i} \) is large, then \( i \) invests more. It is also easy to check that \( y_{i,t} \) decreases over time and reaches zero in the last period. If the pledges are to be decided on again already in the next period, then a party does not invest more than the business-as-usual level, so the additional investment \( (y_{i,T}) \) is zero. The intuition for this is the classical hold-up problem: another unit of technology at the beginning of a new commitment period makes it possible to reduce emission by one unit, forever after, without changing the levels of consumptions or investments. This benefits everyone, not only the party that invests, just as in BAU. So, even with commitments to \( x_{i} > 0 \), the investment levels are the same as in BAU if \( T = 1 \).

### 3.4 Equilibrium Contributions

The previous lemma stated that technology and investment levels 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 function that is quadratic in \( x_{i} \). In fact, the continuation value \( U_{i}(x) \) simplifies to Example E, introduced in Section 2, with \( \alpha, \beta \) and \( \gamma \) functions of \( b, c, k, \delta, \) and \( T \):.

**Lemma 3:** Since every \( y_{i,t} \) is linear in \( x_{i} \), \( i \)'s continuation value, relative to BAU, can be written as in
Example E:

\[ U_i(x) = \sum_{t=1}^{\infty} \delta^t u_{i,t} = \alpha \sum_{j \neq i} x_j - \frac{\beta}{2} x_i^2 + \gamma, \quad \text{where} \]
\[ \alpha = \frac{a}{1 - \delta} \left[ 1 - \delta^T (C_1 L_1^{T-1} + C_2 L_2^{T-1}) \right], \]
\[ \beta = \sum_{t=1}^{T} \delta^{t-1} \left[ b (C_1 L_1^{t-1} + C_2 L_2^{t-1})^2 + c (C_1 L_1^{t-1} [1 - L_1] - C_2 L_2^{t-1} [L_2 - 1])^2 \right]. \]
\[ \gamma = \delta^T U_i(x^*). \]  

Since \( U_i(x) \) is symmetric, Corollary 1 implies that, with pledge-and-review bargaining:

\[ x_i^* = \arg \max_{x_i} U_i(x) + w \sum_{j \neq i} U_j(x) = w (n - 1) \alpha / \beta. \]  

The smaller is \( w \), the smaller are the \( x_i^* \)'s, and the smaller are all the investment levels. Both effects make the parties worse off, relative to a situation in which \( w \) were larger. By combining (12), (E), and substituting in for \( \gamma = \delta U_i(x^*) \), we can see that \( U_i \) increases in \( w \) for every \( w \leq 1 \).

**Proposition 1:** A smaller \( w \) reduces contributions, investments, and therefore payoffs:

\[ U_i(x^*) = \alpha (n - 1) \left[ w (n - 1) \frac{\alpha}{\beta} - \frac{\beta}{2} \left( w (n - 1) \frac{\alpha}{\beta} \right)^2 \right] + \gamma = \frac{\alpha^2 (n - 1)^2}{\beta \left( 1 - \delta^T \right)} w \left( 1 - \frac{w}{2} \right). \]  

**Fact 1.** As explained in Sections 1-2, the Paris Agreement on climate change is associated with pledge-and-review, while the top-down negotiations associated with the Kyoto Protocol can be approximated by the NBS. The smaller \( w \) reduces welfare, according to Proposition 1, and the result therefore rationalizes the criticism mentioned in the Introduction. The following sections show that the picture will be more nuanced when we endogenize participation.

4 Participation

This section endogenizes the coalition size and studies how it depends on the bargaining procedure. The typical way of endogenizing the coalition size is to follow the approach discussed in the Introduction. In this literature, the game begins with a participation stage at which every potential party, \( i \in \{1, ..., n\} \), decides whether or not to participate in the coalition. These decisions are made simultaneously and everyone expects that the participants will negotiate their contribution levels according to the pledge-and-review bargaining game.\(^{25}\) Given the restriction to MPEs, the free-riders will simply follow their

\(^{25}\)In reality, after the treaty is signed, a negotiated treaty must also be ratified by a number of the countries’ governments before it enters into force. It is politically costly to sign but not ratify, and the Vienna Convention on the Law of Treaties (Article 18) verifies that the signature is binding: "A State is obliged to refrain from acts which would defeat the object and purpose of a treaty when...it has signed the treaty..." Therefore, almost every country that signed the Kyoto Protocol (or the Paris Agreement) has also ratified; US being the famous example. Minimum participation thresholds are dealt with in Section 5.
dominant business-as-usual strategy and set $x_i = 0$.

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 out assuming that the participation decision is made once and for all, as in Figure 2, but Section 8 explains that the results are robust and hold also when this assumption is relaxed.

I start by ignoring the constraint $n \leq \pi$; and also a possible minimum participation threshold, $\pi$, but these constraints are extensively discussed in the next section. Since the parties’ payoffs in the climate policy model can be summarized as Example E, I henceforth restrict attention to these payoffs.

### 4.1 Equilibrium Participation

Since the coalition members will contribute more than the level that would maximize their own utility, there is a cost of participating in the coalition. This cost must be smaller than the benefit of participating for a member to be willing to participate. The benefit of participating is that the other participants will take into account (a fraction $\frac{w}{n}$ of) the utility of one additional coalition member.

The payoff for each of the $n$ participants is given by (13). If one of these parties instead free rides, the free-rider’s payoff will be $\alpha (n-1) w (n-2) \alpha/\beta \left(1 - \delta^T\right)$, since each of the other $n-1$ parties will now contribute $w (n-2) \alpha/\beta$ every $T$ period. By comparison, participation is beneficial if:

$$\frac{\alpha^2}{\beta} (n-1)^2 w (1 - w/2) \geq \alpha (n-1) w (n-2) \alpha/\beta \Rightarrow \quad (n-1) w \leq 2.$$  \hfill (14)

The size $n$ cannot be too great since then individual contributions would be so large that free riding would be preferable. For $n$ to be an equilibrium coalition size, (14) must hold for the equilibrium $n$, but it must fail for any larger $n$ (since, otherwise, nonmembers would also like to participate).

**Proposition 2:** The equilibrium coalition size is decreasing in $w$:

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

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

Importantly, $n$ decreases in $w$. According to Proposition 2, $n = 3$ if $w = 1$, as when applying NBS. As discussed in the Introduction, this small-coalition result is well known in the literature, which also discusses the trade-off between ‘narrow and deep’ vs. ‘broad and shallow’ coalitions. With pledge-and-review bargaining, $w$ is small and a coalition member is not expected to contribute a lot. The cost of participation is then small, and participation is attractive for a larger set of $n$’s.

Since the number of participants must be an integer, $n$ is a step-function that decreases in $w$. When comparing pledge-and-review to NBS, we are interested in large rather than small differences in $w$. Thus,
it is not unreasonable to abstract from the fact that \( n \) must be an integer and use the approximation:

\[
n \approx n(w) = 1 + 2/w. \tag{15}
\]

With this approximation, the product \((n(w) - 1)w\) stays constant if \( w \) changes. To see the intuition for this invariance, note that at the equilibrium coalition size, a party is roughly indifferent between whether to participate or free-ride. On the one hand, a constant \((n-1)w\) implies that \( x_i \) is constant and thus the cost of participation remains unchanged when \( w \) changes. On the other hand, the benefit of participating is that each of the \( n-1 \) other parties will increase their \( x_j \)'s by an amount that is proportional to \( w \) if \( i \) participates. This benefit is proportional to \((n-1)w\). Thus, the benefit as well as the cost (and therefore the indifference condition) are unchanged when \( w \) decreases if just \( n \) simultaneously increases so much that the product \((n-1)w\) remains unchanged.

\subsection*{4.2 Implications}

When \((n-1)w\) stays constant as \( w \) is reduced, \( x_i \) also remains constant, and so does every investment level \( y_{i,t} \). Since the individual contributions are invariant in \( w \), while \( n \) is decreasing in \( w \), overall welfare will be larger when \( w \) is small. A participant’s payoff is also larger when \( w \) is small: this is evident when the endogenization of \( n \), as described by (15), is combined with the utility (13). This gives:

\[
U_i^* = \frac{4\alpha^2}{\beta(1 - \delta^T)} \left( \frac{1}{w} - \frac{1}{2} \right). \tag{16}
\]

**Corollary 3:** With endogenous participation, approximated by \( n(w) \), Proposition 1 is reversed: A smaller \( w \) increases aggregate contributions, investments, and welfare.

**Fact 2.** While only 37 countries promised to cut emission in the Kyoto Protocol’s first commitment period, 195 countries have pledged to contribute to the Paris Agreement. This fact is consistent with Proposition 2 since the Paris Agreement is associated with pledge-and-review and thus a smaller \( w \). The popular support is the Paris Agreement’s claim to success, and it can reverse how we may rank the two types of bargaining games, according to Corollary 3.

\section*{5 When to Choose Pledge-and-Review}

With reasonable modifications, as when \( n \) is constrained, the preference over bargaining procedure involves a trade-off. To shed light on when pledge-and-review is preferable, this section compares the participants’ payoffs given a low \( w = \underline{w} \) and a large \( w = \overline{w} > w \). In particular, the comparison may be between NBS \((\overline{w} = 1)\) and P&R \((\underline{w} < 1)\).
5.1 Maximal Participation

The world consists of a finite number of countries, \( n \). In fact, if \( \pi < n(\pi) < n(w) \), where \( n(\cdot) \) is defined by (15), then both bargaining games (characterized by \( w \) or \( \pi \)) ensures full participation. In this case, \( \pi \) is preferable, according to Proposition 1. If, instead, \( n(\pi) < n(w) < \pi \), the upper boundary on \( n \) is nonbinding and \( \pi \) is preferable, according to Corollary 3.

A trade-off arises when \( n(\pi) < n(w) \), since then participation is larger, but individual contributions smaller, when \( w \) is small. In this case, a sufficiently large \( \pi \) is necessary to ensure that a participant’s payoff is largest under \( \pi \).

The exact condition follows when comparing a participant’s utility, as given by equation (13), for the two cases. The payoff is larger when \( w = \pi \) than when \( w = \pi \) if:

\[
\frac{\alpha^2 (\pi - 1)^2}{\beta (1 - \delta^t)} w^2 \left( \frac{1}{w} - \frac{1}{2} \right) > \frac{\alpha^2 (n(\pi) - 1)^2}{\beta (1 - \delta^t)} \pi^2 \left( \frac{1}{\pi} - \frac{1}{2} \right) \Rightarrow \frac{\pi - 1}{n(\pi) - 1} > \Omega, \text{ where}
\]

\[
\Omega \equiv \sqrt{\frac{\pi (1 - \pi/2)}{w (1 - w/2)}} \in \left(1, \frac{\pi}{w}\right).
\]

5.2 Minimum Participation

Countries are more heterogeneous in reality than the model above has permitted. If the willingness to participate varied across countries, the countries that would benefit less from participating would be the marginal ones to participate when \( w \) is reduced. Countries that benefit the most from participating would strictly prefer to participate, and if \( w \) is larger, \( n \) may not decline as fast as predicted by Proposition 2.

There are several ways of capturing this reasoning in the model. Analytically, the simplest way is to assume that a number \( n \) are committed to participate regardless of \( w \). The reason for why these countries are committed can be outside of the model, but one may think off existing international treaties on non-climate issues such as international trade or regulatory politics. To be specific, the European Union member countries cannot easily opt out of an environmental agreement unilaterally.

Alternatively, if the treaty is specifying a minimum participation threshold for the treaty to enter into force, then \( \underline{n} \) may be interpreted as this threshold.\(^{26}\)

The minimum participation level \( \underline{n} \) is relevant only if \( \underline{n} > n(\pi) \). If also \( \underline{n} > n(w) \), the number of participants is always \( \underline{n} \) and then the larger \( w \) is optimal, according to Proposition 1. To isolate the trade-off associated with \( \underline{n} \), suppose \( n(\pi) < \underline{n} < n(w) < \pi \). In this case, only committed parties participates under \( \pi \), while participation under \( w \) is given by \( n(w) \). By comparison, a participant’s payoff can be largest under \( w \) if and only if \( \underline{n} \) is sufficiently small.

\(^{26}\)When the constraint \( n \geq \underline{n} \) is binding, then the \( \underline{n} \) parties must agree unanimously, as assumed in Section 2. When \( n > \underline{n} \), then it feasible to proceed without an unanimous approvement, but if \( n \) countries (are expected to) contribute in equilibrium, then every one of them must prefer the agreement to free-riding. In a pure-strategy equilibrium, one can correctly anticipate which countries will participate and, when \( n \geq \underline{n} \) binds, these are here referred to as the \( \underline{n} \) committed parties.
Participation and participants’ payoffs are strictly decreasing in $w$ only when $n(w) \in (\bar{n}, \bar{n})$.

Again, the exact condition follows when we use the utility function (13) to compare the two cases. The payoff is larger when $w = w$ than when $w = \bar{w}$ if:

\[
\frac{\alpha^2 (n(w) - 1)^2}{\beta (1 - \delta^2)} w^2 \left( \frac{1}{w} - \frac{1}{2} \right) > \frac{\alpha^2 (\bar{n} - 1)^2}{\beta (1 - \delta^2)} \bar{w}^2 \left( \frac{1}{\bar{w}} - \frac{1}{2} \right) \Rightarrow \frac{n(w) - 1}{n - 1} > \Omega.
\]

5.3 The Preferred Bargaining Game

It is possible that both the minimum and the maximum participation levels are of relevance. This happens if $n(\bar{w}) < \bar{n} < \pi < n(w)$. In this case, there is full participation under $w$, but only the committed parties participate under $\bar{w}$. In this situation, $w$ is preferred when $\pi$ is large and $\bar{n}$ is small.

From (13), $w$ gives a higher payoff than $\bar{w}$ if:

\[
\frac{\alpha^2 (n(w) - 1)^2}{\beta (1 - \delta^2)} w^2 \left( \frac{1}{w} - \frac{1}{2} \right) > \frac{\alpha^2 (\bar{n} - 1)^2}{\beta (1 - \delta^2)} \bar{w}^2 \left( \frac{1}{\bar{w}} - \frac{1}{2} \right) \Rightarrow \frac{\bar{n} - 1}{\bar{n} - 1} > \Omega.
\]

The three conditions can be combined in the following way.

**Proposition 3:** Participants prefer pledge-and-review bargaining (i.e., to switch from $w = \bar{w}$ to $w = w < \bar{w}$) if $\pi$ is large while $\bar{n}$ is small. The exact condition is:

\[
\min \left\{ \frac{\pi - 1}{2/w} \right\} > \Omega.
\]
This condition is better illustrated as the solid line in Figure 4. If there is a larger number of potential parties, or if fewer of them are committed to participate when $w$ is large, we move in the direction of the arrow in the figure. Then, the "shallow" agreement ($w$) becomes preferred even though the "deep" agreement was preferred for a smaller number of potential parties, or for a larger number of committed parties.

**Fact 3: From Kyoto to Paris.** One may argue that both these developments (i.e., a larger $\pi$ and a smaller $n$) are in line with changes in world politics over the last few decades. Today we have a larger number of countries that are emerging economies, although in the 1990s they were developing countries that could not be expected to contribute (much) to an international climate change treaty. For the model, this implies that the number of relevant parties, $\pi$, has increased.

During the same period, seven of the original Annex I countries, who initially signed the Kyoto Protocol, announced that they would not make commitments under the Kyoto Protocol’s second commitment period.\textsuperscript{27} These withdrawals may be interpreted as a reduction in the number of committed countries, $n$.\textsuperscript{28} For either (or both) reason, the switch to P&R is thus consistent with the theory of this paper.

To get a sense for exactly when pledge-and-review is preferred, note that the Paris Agreement succeeded in motivating nearly every country in the world to participate. Given that the US has announced that it will withdraw, it may be reasonable to assume that the weight associated with pledge-and-review

\begin{footnote}
\textsuperscript{27}The seven are Belarus, Ukraine, Japan, New Zealand, Russia, Canada, and USA (IPCC 2014:1025: "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."

\textsuperscript{28}If $n$ is instead interpreted as the minimum participation threshold for the agreement to enter into force, then it would be hold constant (equal to 55) for both agreements, according to the two agreement texts.
\end{footnote}
satisfied:
\[ n(w) \approx 195 \iff w \approx \frac{1}{97}. \]

For the sake of illustration, suppose that the Kyoto Protocol can be approximated by the NBS, with \( w = 1 \). When we substitute these numbers into inequality (17), Proposition 3 implies that the Paris Agreement (\( w \)) is preferred to the Kyoto Protocol (\( w \)) if and only if \( n \) is less than 28:

\[ \text{P&R} \succ \text{NBS} \iff n \leq 28. \]

So, with 37 committed parties, as in the 1990s, the Kyoto Agreement is preferred. If the number of committed parties is reduced to 28, the participants prefer pledge-and-review in order to motivate the larger set of countries to participate. Interestingly, the European Union, evidently including the most committed set of countries, consists of exactly 28 countries.

It is straightforward to show that the uncommitted countries prefer to the broad agreement for a smaller set of parameters, and, in the numerical example, only when \( n \leq 20 \).\(^{29}\) The theory can thus rationalize why developing countries preferred to continue with the Kyoto Protocol rather than the P&R system.\(^{30}\) To see the intuition for this, note that the participants can benefit in two ways from a larger \( n \). First, a larger \( n \) might lead to larger aggregate contributions. This benefit is clearly positive if \( n \) and \( \pi \) are not binding \( n(w) \), so that \( n(w) \in (\pi, \bar{\pi}) \), but a larger \( n \) (motivated by a smaller \( w \)) can also lead to less aggregate contributions, when one of the constraints bind (Figure 3). Second, a larger \( n \) implies that the sum of contributions is divided on a larger number of parties. This effect is always positive for the original participants, and it may motivate them to switch to pledge-and-review even in the circumstance in which the first effect, that on total contributions, is negative. In that situation, the switch to pledge-and-review is clearly harmful to all countries that were not participating when \( w = \bar{w} \). Thus, the original set of participants prefer to switch to pledge-and-review too soon, i.e., for a larger set of parameters than the set under which such a switch increases global welfare. Similarly, if it is the new potential members who are pivotal in the decision on treaty design, they will accept pledge-and-review too late or too seldom, relative to the decision that would be optimal if the original members’ payoffs had been taken into account.

\(^{29}\)Countries that participate in the broad but not in the narrow agreement, prefer the broad agreement if and only if:
\[
\min\{\pi - 1, 2/\bar{w}\} > \sqrt{\frac{\max\{\bar{w} (w - 1) \pi, 4/\bar{w} + 2\}}{w (1 - w/2)}}.
\]

\(^{30}\)Bodansky et al. (2017:202) write: "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 Enforcement and Compliance

So far, the analysis has presumed that the pledges are contractible, credible, and complied with. Given the incentive to free ride, discussed in Sections 4 and 5, 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. Since decisions are made simultaneously, a party that "defects" by not contributing will be able to enjoy the benefit from the other participants’ contributions in that period.

Since $i$’s cost is a convex function of $i$’s contribution level, the temptation to defect is largest when the contribution level is large, i.e., when $w$ is large. It is thus intuitive that the compliance constraint is more likely to hold when $w$ is small, as under P&R bargaining.

To illustrate this intuition in a simple way, suppose the parties revert to the noncooperative MPE — the BAU — as soon as one party has defected by contributing less than pledged. The analysis is particularly straightforward when investments are so costly that no-one invests since the dynamic game then boils down to a repeated game. When $c \to \infty$, $y_{i,t} \to 0$, $x^*_i \to w(n-1)/b$, and $U_i(x^*) \to \frac{(n-1)^2 a^2}{b} w \left(1 - \frac{w}{2}\right)$. Party $i$’s payoff when defecting is $aw(n-1)^2 a/b$. By comparison, it is better to comply if:

$$aw(n-1)^2 \frac{a}{b} \leq U_i(x^*) = \frac{(n-1)^2 a^2}{b} w \left(1 - \frac{w}{2}\right) \Rightarrow w \leq 2\delta.$$ 

In general, the dynamic climate change game in Section 3 is somewhat different from the standard cases in the repeated games literature, since past investments influence BAU and thus all future contributions. When the other participants invest $y_{j,t}$, then $j$’s contribution will increase by $y_{j,t}$ from the next period on, and forever thereafter, even if the parties are reverting to the noncooperative MPE. Since every $y_{j,t}$ 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 in the following inequality. This payoff must be smaller than $i$’s equilibrium payoff, written on the right-hand side:

$$a \left( \sum_{j \neq i} x_j + \frac{\delta}{1-\delta} \sum_{j \neq i} y_{j,1} \right) \leq U_i = \frac{a^2(n-1)^2}{b} w \left(1 - \frac{w}{2}\right) \Rightarrow$$

$$w \leq 2 - 2 \left[ 1 - \delta \left( C_1 L_1 + C_2 L_2 \right) \right] \frac{a \left( \frac{1-\delta ^T}{1-\delta} \right) \alpha}{\alpha(1-\delta)} .$$

---

31 A few remarks are in order: The strategies to revert to BAU after a defection are not Markov. Although I here consider history-contingent strategies, it is beyond the scope of this paper to characterize the entire set of subgame-perfect equilibria. On the one hand, it is possible to sustain as SPEs harsher punishments (following defection) than the reversion to BAU: with harsher punishments, a treaty would be self-enforcing for a larger set of circumstances than those derived below. On the other hand, if the parties could renegotiate the punishments, then a treaty would be self-enforcing for a smaller set of parameters.

32 If $c \to \infty$ (or $b/c \to 0$), then $L_1 \to 1$, $L_2 \to 1/\delta$, $C_1 \to 1$, and $C_2 \to 0$, so $y_{i,t} \to 0$ follows from Lemma 2. Also, $\alpha \to \frac{1-\delta ^T}{1-\delta} a$ and $\beta \to \frac{1-\delta ^T}{1-\delta} b$, which can be substituted into (E).
The implication follows when we substitute in for the equilibrium $y_{i,1}$, $x_i$, and $\alpha$ and rewrite. This condition boils down to $w \leq 2\delta$ when $c \to \infty$. In either case, the relevant condition is easier to satisfy when $w$ is small, i.e., if the bargaining procedure is characterized by pledge-and-review rather than the NBS, for example.

Note that $n$ drops out from the inequality, and thus $n$ does not influence whether the bargaining outcome will be self-enforcing. The intuition for this invariance is that both the cost of the individual contribution and the benefit from the others’ contribution are proportional to $(1 - n)^2$. Consequently, the condition is robust to whether $n$ is exogenous (as in Section 3) or endogenous (as in Section 4).

**Proposition 4:** The bargaining outcome is more likely to be self-enforcing if $w$ is small. This result holds whether or not participation is endogenous.

The proof, in the Appendix, permits the punishment to last for any number $t' \leq \infty$ of periods and to be triggered with any probability $p \leq 1$. For every $t'$ and $p$, the incentive constraint will hold if and only if $w$ is sufficiently small, just as above.\(^{33}\)

If $w$ is large and the negotiated commitment levels are not self-enforcing, then the parties must find additional ways of raising the cost of non-compliance. In reality, there are several ways of increasing these costs, since 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.’ Keohane and Oppenheimer (2016) explain 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)."

**Fact 4 (legal status).** Since the Paris Agreement applies pledge-and-review bargaining, where $w$ is smaller, it is more likely that the compliance constraint holds for this agreement without making it legally binding, according to Proposition 4. The pledges are indeed not legally binding under the Paris Agreement, but, as is also consistent with the theory above: "the Kyoto Protocol represents a much harder, more prescriptive approach, including legally binding, quantified emissions limitation targets" (Bodansky and Rajamani, 2017:22). This difference between the two treaties is clearly consistent with Proposition 4.\(^{34}\)

---

\(^{33}\)As proven in the Appendix, the condition is:

$$w \leq 2 - 2 \left[ \frac{1 - \delta \left( C_1 L_1 + C_2 L_2 \right)}{(1 - \delta) \left( 1 - \delta \left( 1 - p + p\delta' \right) \right)} \right] \frac{\alpha \left( 1 - \delta' \right)}{\alpha} \Rightarrow w \leq 2 - 2 \frac{1 - \delta}{1 - \delta \left( 1 - p + p\delta' \right)} \text{ if } c \to \infty.$$

\(^{34}\)Of course, when one can raise the cost of noncompliance by modifying the legal status of the agreement, then countries are likely to comply on the equilibrium path, regardless of the bargaining procedure. As predicted, also for the Kyoto Protocol, the final "thirty-six Kyoto parties were in full compliance with their first commitment period targets" (Bodansky and Rajamani 2017:31).
7 The Review Period Length

The larger is the length of the commitment period, \( T \), the larger are the equilibrium investments at every point in time: this can be seen from Lemma 2. The intuition for this comparative static is the standard hold-up problem: if the next bargaining stage is near in time, then each party invests less because the other parties will expect larger contributions from a party that has invested in the capacity to contribute. The hold-up problem is thus an argument in favor for a longer commitment period. On the other hand, after investments have been made, it is ex post optimal for all the parties to start the pledge-and-review bargaining game soon again, to take advantage of the newly developed technology. The optimal \( T \) trades off the effect on investments with the benefit that newly developed technology can strengthen the commitments sooner when \( T \) is small.\textsuperscript{35}

Despite the fact that there is a trade-off when it comes to deciding on the optimal \( T \), the condition for the optimal \( T \) is independent of \( w \) and \( n \) in the model above. When \( n \) is exogenous, then a party’s payoff is given by equation (13), where both \( \alpha \) and \( \beta \) are functions of \( T \). When \( n \) is instead endogenous, a party’s payoff is given by (16). The optimal choice of \( T \) is the same in both cases:

\[
T^* = \arg \max_T U_i(x^*) = \arg \max_T U_i^* = \arg \max_T \frac{\alpha^2}{\beta (1 - \delta T)}.
\]

where \( \alpha \) and \( \beta \) are functions of \( T \), as described by Lemma 3. If \( c \to \infty \), or \( b/c \to 0 \), investments are irrelevant, and \( \frac{\alpha^2}{\beta (1 - \delta T)} \to \frac{a^2}{b (1 - \delta)} \) is independent of \( T \). In this case, every \( T \) is equally good.

**Proposition 5:** The optimal commitment period length, \( T^* \), is independent of \( n \) and \( w \), and of whether \( n \) and \( w \) are endogenous or exogenous.

**Fact 5.** Given the many differences between the Kyoto Protocol and the Paris Agreement, the two are surprisingly similar when it comes to how frequently the commitments are to be updated. The pledges under the Paris Agreement will be updated every five years, and also the Kyoto Protocol’s first commitment period was five years (2007-2012). It is reasonable that also the second commitment period would have been five years, if the parties had not anticipated that a new global treaty would be effective from 2020.\textsuperscript{36} This similarity is consistent with Proposition 5, stating that there is no reason to change the commitment period length, despite the many other differences between two treaties.

\textsuperscript{35}The combined trade-off is new to the literature: The hold-up problem associated with a smaller \( T \) is already recognized (see, e.g., Harstad, 2016, or, with renegotiation: Rey and Salanie, 1990; Harstad, 2012). Harris and Holmstrom (1987) observed that a smaller \( T \) may be necessary in order to update the terms of the contract.

\textsuperscript{36}According to Bodansky et al. (2017:203), 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 (p. 205), "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."
The model above is simple and stylized and yet able to rationalize facts 1-5, discussed above. This rationalization is quite robust in that it continues to hold for a number of model modifications. This section explains (and the online appendix proves) that Propositions 1-5 hold also if the parties negotiate investment levels or emission taxes (or both) instead of (or, in addition to) the $x_i$’s. The $x_i$’s can also be time-dependent, and the investment levels might be decided on by profit-maximizing firms, without changing the propositions. The results are also quite robust to changes in the timing.

(i) Pledging to invest. Some of the INDC’s in the Paris Agreement specify national targets for renewable energy. This possibility can be captured in the model by letting parties decide on the $y_{i,t}$’s instead of the emission cuts. As discussed in the online appendix, it is straightforward to analyze this scenario: when the $y_{i,t}$’s, but not the $x_{i,t}$’s, are pinned down, then $i$’s consumption will satisfy $b(x_{i,t} - Y_{i,t}) = 0$, just as in BAU. If the investment level must be time-independent ($y_i$) over a commitment period, then $i$’s continuation value can be written as in Example (E), where $x_i$ is replaced by $y_i$, although the definition of $\alpha$ and $\beta$ will be different. The proofs of Propositions 1-5 are thus similar to before.

In fact, $i$’s continuation value will be separable in the $y_t$’s, where $y_t = (y_{1,t}, ..., y_{n,t})$. Consequently, we can apply Corollary 1 when the parties negotiate $y_t$, while keeping fixed the investment levels for the other periods. Corollary 1 will imply that the pledge-and-review outcome for $y_{i,t}$ is:

$$y_{i,t}^* = (n - 1) w \frac{\delta c}{1 - \delta}.$$  

Since 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 equally good, regardless of the level of $n$ and/or $w$.

(ii) Pledging on emission taxes. It is also straightforward to allow the parties to pledge on domestic emission taxes, instead of emission cuts. With an emission tax $\tau_{i,t}$, it is natural that the consumption of fossil fuel is given by the condition in which the marginal benefit of consuming more equals the tax. This implies $b(x_{i,t} - Y_{i,t}) = \tau_{i,t}$. However, when the parties are free to decide on their investment levels, they will invest just as in BAU, so $y_{i,t} = 0$. If the emission tax level must be time-independent ($\tau_i$) over a commitment period, then $i$’s continuation value can be written as in Example (E), where $x_i$ is replaced by $\tau_i$, although the definition of $\alpha$ and $\beta$ will be different. Again, the proofs of Propositions 1-5 are similar to before.

In fact, $i$’s continuation value will be separable in the $\tau_t$’s, where $\tau_t = (\tau_{1,t}, ..., \tau_{n,t})$. Consequently, we can apply Corollary 1 when the parties negotiate $\tau_t$, while keeping fixed the investment levels for the

---

37For example, China has pledged to increase the share of non-fossil fuels in primary energy consumption to around 20 percent, while India is pledging to achieve about 40 percent cumulative electric power installed capacity from non-fossil fuel based energy resources by 2030. See http://cait.wri.org/indc/#/ for a recent overview.
other periods. Corollary 1 will imply that the pledge-and-review outcome for $\tau_{i,t}$ is:

$$\tau_{i,t}^* = (n - 1) wa.$$  \hspace{1cm} (20)

Since this $\tau_{i,t}$ is time-independent, there is no loss for the 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 level of $n$ and/or $w$.

As a side-remark, it is worth noting that the choice of instrument (i.e., whether the parties should negotiate $x_i$'s, $y_i$'s, or $\tau_i$'s) is also independent of $n$ and $w$. As proven in the online appendix, negotiating investment levels are better for all the parties than negotiating emission taxes if and only if investments are inexpensive and the future is important: 38

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

(iii) Pledging on investment levels and emission taxes. Party $i$'s continuation value is separable in the $y_i$'s and the $\tau_i$'s, it can be shown. Thus, Corollary 1 can be applied for each instrument separably, while keeping the other fixed. With this procedure, the outcome is given by the combination of the two above, (19) and (20). When both investment levels and emission taxes can be negotiated, we have a "complete contract" since, given the negotiated investment levels (and thus the $Y_i,t$'s), the emission taxes pin down the contribution levels. Also in this case, continuation values can be written as in Example E, and the earlier proofs apply.

(iv) Pledging on investment levels and contribution levels. Once the investment levels (and thus the $Y_i,t$'s) are pinned down, negotiating $\tau_{i,t} = b(x_{i,t} - Y_{i,t})$ is equivalent to negotiating $x_{i,t}$. Thus, scenario (iii) gives the same outcome as if the parties can negotiate every investment level as well as every contribution level. As before, the choice of $T$ is here irrelevant, regardless of the level of $n$ and/or $w$. 39

(v) Time-dependent contribution levels. In scenario (iv), one best choice of $T$ is $T = \infty$. For $T = \infty$, it is actually irrelevant that the parties have negotiated the investment levels in addition to the contribution levels. The irrelevance follows because, once the $x_{i,t}$'s are given for all future, there is no externality associated with the $y_{i,t}$'s, and, hence, every party will have the right incentive to invest optimally, without any need to negotiate $y_{i,t}$.

As is shown in the online appendix, the equilibrium time-dependent contribution level is given by:

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

38 The comparison to the situation in which the $x_i$'s are negotiated is more complex, however.

39 If the parties can negotiate time-independent $x_j$'s and $y_j$'s, which must stay constant over 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) as well as scenario (iii)), while $T > 1$ would be less efficient. In contrast to the discussion on the optimal $T$, in the previous section, there is no need to have a large $T$ when the first-period investment level can be negotiated, since agreeing on $y_{i,1}$ circumvent the hold-up problem.
With this, party $i$ will voluntarily invest as in (19), ensuring that the marginal benefit from consuming (and from cutting emissions) is $b (x_{i,t}^* - ty_{i,t}) = (n - 1) wa$, which equals $\tau_{i,t}^*$, derived above.

In this situation, it is clear that parties are strictly better off with $T = \infty$ than with $T < \infty$, since, with any finite $T$, equilibrium $y_{i,t}$'s are lower (and less efficient) than the $y_{i,t}$ that would follow in scenario (iv), which coincides with the equilibrium $y_{i,t}$'s when $T = \infty$. When referring to the trade-off discussed in the previous section, there is no reason to reduce $T$ in order to update the pledges, when the pledges can be time-dependent. Given the hold-up problem, it is therefore optimal with $T = \infty$ when contribution levels can be time-dependent.

(vi) **Firms invest.** All the three scenarios (iii)-(v) implement the complete contract outcome, i.e., as when every $y_{i,t}$ and $x_{i,t}$ are negotiated (according to pledge-and-review). The same outcome can be achieved if the parties negotiate $x_{i,t}$ at time $t$, for $T = 1$, while letting firms invest. If so, the pledged $x_{i,t}$ will satisfy $b (x_{i,t} - Y_{i,t}) = (n - 1) wa$, which also characterize the marginal willingness to pay for another unit of $Y_{i,t}$ at time $t$. Thus, the present-discounted value of another invested unit is $\delta (n - 1) wa / (1 - \delta)$, while the marginal investment cost is $cy_{i,t}$. The two are equalized when profit-maximizing price-taking firms decide on $y_{i,t}$ and, then, the result is $y_{i,t}^*$, 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 parties, are not discouraged by the nations’ hold-up problem when new pledges are negotiated.\textsuperscript{40}

(vii). **The timing of $T$.** Proposition 5 showed that everyone agreed on the choice of $T$ and this choice was independent of $n$ and $w$. Thus, the choice of $T$ remains the same also if $T$ is decided on after $w$ is decided on, whether or not $T$ is decided before the participation stage or after the participation stage.\textsuperscript{41} The timing of $T$ does not influence the equilibrium level of $n$ or the preference over $w$, either. In fact, both $n$ and $w$ are independent of $T$, it has been shown.

(viii). **Multiple participation stages.** The results above are identical if there is a participation stage before the pledges are negotiated at the beginning of every commitment period (i.e., every $T$ period). In a Markov-perfect equilibrium, the payoff party $i$ can expect after period $T$ will be independent of whether it participates at $t = 1$. Participating in the first commitment period is then attractive if and only if (14) holds, just as before, because the term $\left(1 - \delta^T\right)$ is, in any case, cancelling from the expression. The identity of the $n$ participants is also the same in every commitment period (in a Markov equilibrium), implying that every participant’s equilibrium continuation value is given by (16). Thus, the proofs of Propositions 2-5 continue to hold.

\textsuperscript{40}Of course, if each government can regulate (subsidize/tax) the firms’ investments, then the government can implement its preferred choice of investment, as described in the previous sections, and then even the exact equations above stay unchanged.

\textsuperscript{41}If $T$ is decided on before the participation stage, I assume that $T$ is set such as to maximize the participants’ payoffs (for example because $T$ is decided on by the committed parties, referred to in Section 6). Nonparticipants may have different preferences over $T$.\textsuperscript{29}
(ix). **Multiple bargaining-choice stages.** Propositions 2-5 also hold if \( w \), as well as \( n \), are endogenously chosen at the beginning of every commitment period, for the same reasons as in scenario (viii). In fact, if some parameters (such as \( n_2 \) and/or \( \pi \)) change every \( T \) period, then Propositions 2-5 characterize the outcome, and Proposition 3 characterizes the optimal bargaining procedure, for every commitment period, independently of the parameter values after period \( T \). This generalization implies that Proposition 3 can indeed rationalize a change from one procedure to another, when \( n_2 \) and/or \( \pi \) change over time.\(^{42}\)

(x). **Limited punishments.** When the self-enforcement constraint was discussed, Proposition 4 relied on the assumption that if one party defected, then all parties play BAU forever after. On the one hand, one may argue that it is optimistic to assume that a defection can be observed with probability \( p = 1 \). On the other hand, one may also argue that, if cooperation has broken down, then the parties have an incentive to renegotiate and start cooperating again. To capture these concerns to some extent, the proof of Proposition permits defection to be punished with BAU for \( t'_0 \leq \infty \) periods with probability \( p \leq 1 \) (while, with probability \( 1 - p \), there is no punishments, as if the defection was not observed). Although a smaller \( p \) or \( t'_0 \) makes it more difficult to satisfy the incentive constraint, Proposition 4 continues to hold for every \( p \) and \( t'_0 \).

These generalizations can be summarized in the following proposition, detailed and proven in the online appendix.

**Proposition 6:** Propositions 1-5 continue to hold if the parties (P&f/R) bargain:

(i) investment levels instead of \( x \);

(ii) emission taxes instead of \( x \);

(iii) investment levels and emission taxes instead of \( x \);

(iv) investment levels and \( x \) instead of only \( x \);

(v) a time profile \( \{x_t\}_{t=1}^{\infty} \) instead of a time-independent \( x \);

(vi) \( x \), while profit-maximizing price-taking firms invest;

(vii) \( T \), after the \( n \)-stage, or before \( n \) but after the \( w \)-stage.

(viii) Propositions 2-5 continue to hold if there is a participation stage before every commitment period;

(ix) Propositions 3-5 continue to hold if both \( w \) and \( n \) are decided on in every period.

(x) Proposition 4 holds if defection leads to BAU for \( t'_0 \leq \infty \) periods with probability \( p \in (0,1] \).

Needless to say, the choice of \( T \), for example, may depend on many things that are outside of this model, such as policy makers’ ability to commit to the distance future, or the ability to predict the optimal level of contributions many years in advance. Propositions 1-4 have therefore been derived for any fixed \( T \), and they hold regardless of the choice of \( T \).

\(^{42}\)It would be more complicated, however, if parameters varied between every period.
In Harstad (2016), relying on the NBS, shocks on the marginal environmental harm accumulated over time. The shocks make it difficult to predict the optimal pledge many periods in advance, and they motivate a smaller $T$, while the hold-up problem, mentioned above, motivates a larger $T$. In Battaglini and Harstad (2016), $n$ and thereafter $T$ follow endogenously before every commitment period. Then, the countries that have decided to participate may prefer a small $T$ if $n$ is small, since the small $T$ facilitates the admission of new participants sooner. Since the small $T$ is also leading to the hold-up problem, countries are motivated to participate in order to guarantee a larger $T$. I abstract from these extensions here since they would overlap with the earlier papers and since they are not necessary to rationalize the five facts on how the Paris Agreement compares to the Kyoto Protocol. On the contrary, if the modified model does predict that $T$ should be a function of $w$ and/or $n$, or that any of the other propositions will change, then the modified models are not supported by facts 1-5.

9 Concluding Remarks

This paper presents a model and an analysis of pledge-and-review bargaining. The novelty of this bargaining game is that each party proposes how much to contribute individually – unconditional on what other parties pledge – before the parties agree to the vector of pledges. If there is some uncertainty regarding what other parties are willing to accept, for example due to shocks in the short-term discount rate, then contributions will be larger if there is a substantial variance in these shocks. In equilibrium, each party’s contribution level is as described by an asymmetric Nash Bargaining Solution, where the weights on others’ payoffs reflect the distribution and correlation of shocks. Since the weights vary from pledge to pledge, the collection of pledges is not Pareto optimal.

The pledge-and-review bargaining game has been associated with the 2015 Paris Agreement on climate change, and it makes the agreement rather different from the top-down approach that characterized the Kyoto Protocol of 1997. The analysis uncovered that the (1) difference in bargaining procedure can rationalize four other facts on how the Paris Agreement differs from the Kyoto Protocol: (2) Since pledge-and-review bargaining is not very demanding, it is attractive for a larger number of participants. This can explain why many more countries took on emission cuts in Paris than in Kyoto. (3) Since raising participation is the main benefit of pledge-and-review, it is preferably if and only if there is a large number of relevant players. Although the Kyoto Protocol’s top-down procedure was chosen in the 1990s, this logic can explain why pledge-and-review was preferred in the 2010s, after several developing economies had become emerging economies. (4) Since pledge-and-review is not very demanding, the equilibrium pledges are more likely to be self-enforcing. This rationalizes why the Kyoto Protocol’s emission cuts had to be legally binding, while they are not for the Paris Agreement. (5) Despite all these differences, the theory is consistent with the fact that the length of the commitment period is the same for the two types of agreements.
Although the paper restricts attention to a positive analysis, the reader may instinctively search for normative lessons. Pledge-and-review bargaining might not be as inadequate as it at first appears to be; it can actually be preferable to the alternatives when participation is endogenous. However, if participation can be encouraged by other means, then a more demanding conditional-offer bargaining game becomes preferable. The conclusion is thus that the benefit of offering "club benefits" (such as the lower tariffs in Nordhaus, 2015) to coalition members is not, at the end of the day, that participation will increase, but that the coalition can move towards a more ambitious bargaining procedure without fearing that participation will fall by too much.
10 Appendix

Proof of Theorem 1.

As advertised in Section 2, the following generalization of Theorem 1 is here proven without the additional assumptions $\partial U_i (\cdot) / \partial x_i < 0$ for $x_i > 0$, and $\partial U_j (\cdot) / \partial x_i > 0$, $j \neq i$.

**Theorem 1**. If $x^*$ is a nontrivial stationary SPE in which $U_i (x^*) > 0 \forall i$, then, for every $i \in N$, we have:

(a) if $\frac{\partial U_i (x^*)}{\partial x_i} \leq 0$,

$$- \frac{\partial U_i (x^*)}{\partial x_i} \leq \sum_{j \neq i} \max \left\{ 0, \frac{\partial U_j (x^*)}{\partial x_i} \right\} f_j (0) E (\theta_{i,t} | \theta_{j,t} = 0) \rho_j \Delta U_i (x^*);$$

(b) if $\frac{\partial U_i (x^*)}{\partial x_i} > 0$,

$$\frac{\partial U_i (x^*)}{\partial x_i} \leq \sum_{j \neq i} \max \left\{ 0, -\frac{\partial U_j (x^*)}{\partial x_i} \right\} f_j (0) E (\theta_{i,t} | \theta_{j,t} = 0) \rho_j \Delta U_i (x^*).$$

Note that with the additional assumptions $\partial U_i (\cdot) / \partial x_i < 0$ for $x_i > 0$, and $\partial U_j (\cdot) / \partial x_i > 0$, $j \neq i$, the first-order condition of (3) is equivalent to (21).

(a) First, note that in any stationary SPE we must have $U_i (x^*) \geq 0 \forall i$, since otherwise a party with $U_i (x^*) < 0$ would reject $x^*$ in order to obtain the default payoff, normalized to zero. We will search for nontrivial equilibria in which $U_i (x^*) > 0 \forall i$.

A stationary equilibrium $x^*$, such that $U_j (x^*) > 0 \forall j$, is accepted with probability 1 when $\rho_{j,t} \geq 0$. Therefore, $i$ will never offer $x_i > x_i^*$ when $\frac{\partial U_i (x^*)}{\partial x_i} \leq 0$, so to check when $x^*$ is an equilibrium, it is sufficient to consider a deviation by $i$, $x^*$, such that $x_i^* < x_i^*$ while $x_j^* = x_j^*$, $j \neq i$.

**Acceptable offers.** Let $p (x^*; x^*)$ be the probability that at least one $j \neq i$ rejects $x^*$, and $p_{-j} (x^*; x^*)$ the probability that at least one party other than $j$ and $i$ rejects $x^*$.

Since party $j$'s discount factor is $\delta_{j,t}^d = 1 - \rho_{j,t} \Delta = 1 - \theta_{j,t} \rho_j \Delta$, $j \neq i$ rejects $x^*$ iff:

$$\left( 1 - \rho_{j,t} \Delta \right) U_j (x^*) + p_{-j} (x^*) \left( 1 - \rho_{j,t} \Delta \right) U_j (x^*) < (1 - \rho_{j,t} \Delta) U_j (x^*) \Rightarrow$$

$$\theta_{j,t} < \tilde{\theta}_j (x^*) \equiv \max \left\{ 0, \frac{U_j (x^*) - U_j (x^*)}{\rho_j \Delta U_j (x^*)} \right\}.$$

When the joint pdf of shocks $\theta_t = (\theta_{1,t}, ..., \theta_{n,t})$ is represented by $f (\theta_t)$, the probability that every $j \neq i$ accepts $x^*$ can be written as:

$$1 - p (x^*) = G \left( \tilde{\theta}_1 (x^*), ..., \tilde{\theta}_{i-1} (x^*), \tilde{\theta}_{i+1} (x^*), ..., \tilde{\theta}_n (x^*) \right)$$

$$= \int_0^{\tilde{\theta}_i} \left[ \int_{\tilde{\theta}_i (x^*)}^{\tilde{\theta}_{i+1} (x^*)} \cdots \int_{\tilde{\theta}_{i-1} (x^*)}^{\tilde{\theta}_{i+1} (x^*)} \cdots \int_{\tilde{\theta}_n (x^*)}^{\tilde{\theta}_{n+1} (x^*)} f (\theta_t) d\theta_{-i,t} \right] d\theta_i,$$

which is a function of $n - 1$ thresholds. By taking the derivative w.r.t. $x_i^*$ and using the chain rule,

$$- \frac{\partial p (x^*)}{\partial x_i} = \sum_{j \neq i} \max \left\{ 0, \frac{\partial U_j (x^*)}{\partial x_i} \right\} G_j \left( \tilde{\theta}_1 (x^*), ..., \tilde{\theta}_{i-1} (x^*), \tilde{\theta}_{i+1} (x^*), ..., \tilde{\theta}_n (x^*) \right),$$

(22)
and, at the equilibrium, \( x^i = x^* \),
\[
\frac{\partial p(x^*)}{\partial x_i} = \sum_{j \neq i} \max \left\{ 0, \frac{\partial U_j(x^*)}{\partial x_i} - \frac{\rho_j \Delta U_j(x^*)}{\partial x_i} \right\} G_j(0) = -\sum_{j \neq i} \max \left\{ 0, \frac{\partial U_j(x^*)}{\partial x_i} - \frac{\rho_j \Delta U_j(x^*)}{\partial x_i} \right\} f_j(0), \tag{23}
\]
where, as written in the text already, \( f_j(0) \) is the marginal distribution of \( \theta_{j,t} \) at \( \theta_{j,t} = 0 \).

**Equilibrium offers.** When proposing \( x_i \), party \( i \)'s problem is to choose \( x_i \leq x_i^* \) so as to maximize
\[
(1 - p(x^*)) U_i(x^*) + p(x^*) \left( 1 - E_{i,t} \theta_i^R \right) U_i(x^*), \tag{24}
\]
where \( E_{i,t} \theta_i^R \) is the expected \( \theta_{i,t} \) conditional on being rejected (this will be more precise below).

To derive the first-order condition w.r.t. \( x_i \), suppose \( i \) considers a small (marginal) reduction in \( x_i \) relative to \( x_i^* \), given by \( dx_i = x_i^* - x_i > 0 \). If accepted, this gives \( i \) utility \( U_i(x^i) \approx U_i(x^*) + dx_i \partial U_i(x^*)/\partial x_i > U_i(x^*) \), but it is rejected with probability
\[
\frac{\partial p(x^*)}{\partial x_i} dx_i = -\sum_{j \neq i} \max \left\{ 0, \frac{\partial U_j(x^*)}{\partial x_i} - \frac{\rho_j \Delta U_j(x^*)}{\partial x_i} \right\} dx_i f_j(0),
\]
where each of the \( n - 1 \) terms represents the probability that \( \theta_{j,t} \) is so small that \( j \) rejects if \( x_i \) is modified by \( dx_i \), i.e., \( \Pr \left( \theta_{j,t} \leq \tilde{\theta}_j \right) = \frac{\partial U_j(x^*)/\partial x_i}{\rho_j \Delta U_j(x^*)} | dx_i | \). Naturally, the probability that more than one party has such a small shock vanishes when \( | dx_i | \rightarrow 0 \) since \( f \) is assumed to have no mass point.

In combination, the reduction in \( x_i \) is not beneficial to \( i \) iff:
\[
\left( 1 - \frac{\partial p(x^*)}{\partial x_i} dx_i \right) \left( U_i(x^*) + dx_i \frac{\partial U_i(x^*)}{\partial x_i} \right) - \sum_{j \neq i} \max \left\{ 0, \frac{\partial U_j(x^*)}{\partial x_i} - \frac{\rho_j \Delta U_j(x^*)}{\partial x_i} \right\} dx_i f_j(0) U_i(x^*) \left( 1 - E(\theta_{i,t} | \theta_{j,t} \leq \tilde{\theta}_j) \rho_i \Delta \right) \leq U_i(x^*),
\]
where (23) has been substituted into the second line of (25), and where \( E(\theta_{i,t} | \theta_{j,t} \leq \tilde{\theta}_j) \) follows from Bayes' rule:
\[
E(\theta_{i,t} | \theta_{j,t} \leq \tilde{\theta}_j) = \frac{\tilde{\theta}_j \int_{\Theta_{-j}} \theta_{i,t} f(\theta_{i,t}) d\theta_{i,t}}{\int_{\Theta_{-j}} f(\theta_{i,t}) d\theta_{i,t}}, \quad \text{and} \quad E(\theta_{i,t} | \theta_{j,t} = 0) = \lim_{d\theta_{i,t} \rightarrow 0} \frac{\tilde{\theta}_j \int_{\Theta_{-j}} \theta_{i,t} f(\theta_{i,t}) d\theta_{i,t}}{\int_{\Theta_{-j}} f(\theta_{i,t}) d\theta_{i,t}},
\]
and, as already defined, \( \Theta_{-j} = \prod_{k \neq j} [0, \tilde{\theta}_k] \) and \( \tilde{\theta}_j = \frac{\partial U_j(x^*)/\partial x_i}{\rho_j \Delta U_j(x^*)} | dx_i | \).

When both sides of (25) are divided by \( | dx_i | \) and \( dx_i \uparrow 0 \), (25) can be rewritten as the first-order condition (21).

The proof of part (b) is analogous and thus omitted. **QED**

**Proof of Theorem 2.**

A continuum of \( x^* \)'s can satisfy the equilibrium condition in Theorem 1, since it is not necessary for \( i \) to improve an offer relative to \( x^* \) when \( p(x^*) = 0 \). The idea of the Small Trembles is to introduce trembles such that \( p(x^*) > 0 \) and thus we must check that \( i \) cannot benefit from marginally increasing or decreasing \( x_i^* \) from \( x_i^* \) to reduce \( p(x^*) \). With the Small Trembles, \( i \) will strictly benefit from \( dx_i > 0 \) when (3) is strict, and thus it must hold with equality at \( x^* \).
The vector $\epsilon_i$ is i.i.d. over time according to some cdf, $H(\cdot)$. (For simplicity, I omit the superscript $k$.) When $j$ considers whether to accept $U_j \left( x^i + \epsilon_i \right)$, $j$ faces the continuation value $V_j \left( x^* \right)$ by rejecting, where $V_j \left( x^* \right)$ takes into account that $x^*$ may be rejected in the future (if the future $\epsilon_{i,t}$’s are sufficiently small).

The shocks, combined with the possibility to reject, imply that $V_j \left( x^* \right) > 0$ even if $U_j \left( x^* \right) = 0$, so there is no need to assume $U_j \left( x^* \right) > 0 \forall j$.

With this, party $j \neq i$ rejects $x^i$ if and only if:

$$
(1 - p_j \left( x^i \right)) U_j \left( x^i + \epsilon_i \right) + p_j \left( x^i \right) \left( 1 - \rho_j \Delta \right) V_j \left( x^* \right) \Rightarrow 1 - \rho_j \Delta \Rightarrow \frac{U_j \left( x^i + \epsilon_i \right) - U_j \left( x^i + \epsilon_i \right)}{\rho_j \Delta V_j \left( x^* \right)}.
$$

So, the probability that every $j \neq i$ accepts is:

$$
1 - p \left( x^i \right) = \int_{\epsilon} G \left( \bar{\theta}_1 \left( x^i \right), ..., \bar{\theta}_{i-1} \left( x^i \right), \bar{\theta}_{i+1} \left( x^i \right), ..., \bar{\theta}_n \left( x^i \right) \right) dH \left( \epsilon \right)
$$

$$
= \int_{\epsilon} \int_0^{\bar{\theta}_i} \int_{\bar{\theta}_{i-1} \left( x^i \right)}^{\bar{\theta}_{i+1} \left( x^i \right)} \int_{\bar{\theta}_n \left( x^i \right)}^{\bar{\theta}_n \left( x^i \right)} f \left( \theta_i \right) d\theta_i dH \left( \epsilon \right) \Rightarrow
$$

$$
- \frac{\partial p \left( x^i \right)}{\partial x_i} = \int_{\epsilon} \sum_{j \neq i} - \frac{\partial U_j \left( x^i + \epsilon \right)}{\rho_j \Delta V_j \left( x^* \right)} G_j \left( \bar{\theta}_1 \left( x^i \right), ..., \bar{\theta}_{i-1} \left( x^i \right), \bar{\theta}_{i+1} \left( x^i \right), ..., \bar{\theta}_n \left( x^i \right) \right) dH \left( \epsilon \right).
$$

The condition under which $i$ does not benefit from a marginal change $dx_i$ is given by the analogously modified (25),

$$
U_i \left( x^* \right) + \sum_j \frac{\partial U_j \left( x^* \right)}{\rho_j \Delta U_j \left( x^* \right)} f_j \left( 0 \right) E \left( \theta_i \mid \theta_j = 0 \right) \rho_i \Delta U_i \left( x^* \right) = 0,
$$

(26)

43It will be the combination of the $\epsilon_{i,t}$’s and the $\theta_{j,t}$’s that determines whether $j$ rejects $x^*$: let $\Phi_A \left( x^* \right)$ be the set of $\epsilon_{i,t}$’s and $\theta_{j,t}$’s such that every $j$ accepts $x^*$, while $\Phi_R \left( x^* \right)$ is the complementary set. We then have $p \left( x^* \right) = \Pr \left( \left( \epsilon, \theta \right) \in \Phi_R \left( x^* \right) \right)$ and:

$$
V_j \left( x^* \right) = \int_{\epsilon} \sum_{(\epsilon, \theta) \in \Phi_A \left( x^* \right)} \left( 1 - p \left( x^* \right) \right) U_j \left( x^* + \epsilon \right) + p \left( x^* \right) V_j \left( x^* \right) E_{\epsilon_{i,t} \left( \epsilon, \theta \right) \in \Phi_A \left( x^* \right)} \left( 1 - \rho_j \Delta \right)
$$

where the two expectations are taken over the set of parameters leading to acceptance vs. rejections, respectively.

44The modified version of (25) can be written as:

$$
\left( 1 - p \left( x^* \right) \right) \frac{\partial p \left( x^* \right)}{\partial x_i} dR_{\epsilon_{i,t} \left( x^* \right)} \int_{\epsilon} \sum_{(\epsilon, \theta) \in \Phi_A \left( x^* \right)} U_i \left( x^* + \epsilon \right) + \frac{\partial U_i \left( x^* + \epsilon \right)}{\partial x_i} dR_{\epsilon_{i,t} \left( x^* \right)} \int_0^{\bar{\theta}_i} \int_{\bar{\theta}_{i-1} \left( x^* \right)}^{\bar{\theta}_{i+1} \left( x^* \right)} \int_{\bar{\theta}_n \left( x^* \right)}^{\bar{\theta}_n \left( x^* \right)} f \left( \theta_i \right) d\theta_i dH \left( \epsilon \right) =
$$

$$
\int_{\epsilon} \sum_{j \neq i} \left[ \frac{\partial U_j \left( x^* + \epsilon \right)}{\rho_j \Delta V_j \left( x^* \right)} dR_{\epsilon_{i,t} \left( x^* \right)} \right] \int_0^{\bar{\theta}_i} \int_{\bar{\theta}_{i-1} \left( x^* \right)}^{\bar{\theta}_{i+1} \left( x^* \right)} \int_{\bar{\theta}_n \left( x^* \right)}^{\bar{\theta}_n \left( x^* \right)} f \left( \theta_i \right) d\theta_i dH \left( \epsilon \right).
$$

$$
U_i \left( x^* \right) E_{\epsilon_{i,t} \left( x^* \right) \in \Phi_R \left( x^* \right)} \left( 1 - \theta_j \rho_j \Delta \right) \leq U_i \left( x^* \right).
$$
which is the first-order condition of

\[
\arg \max_{x_i} \prod_{j \in N} \left( U_j \left( x_i, x_i^* \right) \right)^{w_i^j},
\]

when \( \frac{w_i^j}{w_i^j} = \frac{\theta x f_j(0)}{\rho_j f_j(0)} E(\theta_{i,t} | \theta, \theta_{i,t} = 0), \forall j \neq i. \) QED

**Proof of Corollary 2.** With (5), a binding (3) implies:

\[
x_i^* = \arg \max_{x_i} \prod_{j \in N} \left( v_j \left( x_j \right) P \left( x \right) \right)^{w_i^j} = \arg \max_{x_i} v_i \left( x_i \right) P \left( x \right) \sum_j w_j^i / w_i^j
\]

\[
= \arg \max_{x_i} v_i \left( x_i \right) w_i^j / \sum_j w_j^i \prod_{j \in N} v_j (x_j) w_j^i / \sum_k w_k^i \prod_{j \in N} v_j (x_j) w_j^i / \sum_k w_k^i P \left( x \right), \text{ so}
\]

\[
x^* = \arg \max_{x} \prod_{j \in N} v_j (x_j) w_j^i / \sum_k w_k^i P \left( x \right),
\]

which also can be written as (6), given the definitions \( \theta_i \) and \( \omega. \)

The first-order conditions for (6) as well as for (7) require that \( \theta_i v_i' \left( x_i \right) / v_i \left( x_i \right) P_i' \left( x \right) \) is the same for every \( i. \) This pins down the relationship between every pair \( \left( x_i, x_j \right), \) given \( x. \) Thus, the two solutions are equivalent when also the \( P \left( x \right)'s \) are the same. QED

**Proof of Lemma 1.** I will first derive \( U_{i,t}^B. \) When we substitute in for \( \tilde{u}_{i,t}, \tilde{x}_{i,t}^{BAU}, \) and \( \tilde{y}_{i,t}^{BAU} \) into \( U_{i,t}^B = \sum_{\tau=t}^{\infty} \delta^{\tau-t} \tilde{u}_{i,\tau}, \) we can rewrite \( U_{i,t}^B \) as:

\[
U_{i,t}^B = \sum_{\tau=t}^{\infty} \delta^{\tau-t} \left[ a \sum_j (\tilde{x}_{j,\tau} - \frac{b}{2} (\tilde{y}_{i,\tau} - \tilde{Y}_{i,\tau})^2 - \frac{c}{2} \tilde{y}_{i,\tau}^2) \right]
\]

\[
= \sum_{\tau=t}^{\infty} \delta^{\tau-t} \left[ a \sum_j (\tilde{Y}_{j,\tau} + \frac{a}{b}) - \frac{b}{2} \left( \frac{a}{b} \right)^2 - \frac{c}{2} \left( \frac{a}{1 - \delta c} \right)^2 \right]
\]

\[
= \frac{a}{1 - \delta} \sum_j \tilde{Y}_{j,t} + a \sum_j \sum_{\tau=t}^{\infty} \delta^{\tau-t} \frac{1}{1 - \delta} \tilde{y}_{j,\tau} + \sum_{\tau=t}^{\infty} \delta^{\tau-t} \left[ \left( n - \frac{1}{2} \right) \frac{a^2}{b} - \frac{c}{2} \left( \frac{\delta}{1 - \delta c} \right)^2 \right]
\]

\[
= \frac{a}{1 - \delta} \sum_j \tilde{Y}_{j,t} + a \sum_j \sum_{\tau=t}^{\infty} \delta^{\tau-t} \frac{1}{1 - \delta} \left( \frac{\delta}{1 - \delta c} \right) + \sum_{\tau=t}^{\infty} \delta^{\tau-t} \left[ \left( n - \frac{1}{2} \right) \frac{a^2}{b} - \frac{c}{2} \left( \frac{\delta}{1 - \delta c} \right)^2 \right]
\]

\[
= \frac{a}{1 - \delta} \sum_j \tilde{Y}_{j,t} + \frac{a^2}{1 - \delta} \left( n - \frac{1}{2} \right) \left( \frac{1}{b} + \frac{1}{c} \left( \frac{\delta}{1 - \delta} \right)^2 \right).
\]

Similarly, the BAU payoff at time \( T + 1 \) can be written as:

\[
U_{i,T+1}^B = \frac{a}{1 - \delta} \sum_j \left( \tilde{Y}_{j,T+1}^{BAU} + \tilde{y}_{j,T+1} \right) + \frac{a^2}{1 - \delta} \left( n - \frac{1}{2} \right) \left( \frac{1}{b} + \frac{1}{c} \left( \frac{\delta}{1 - \delta} \right)^2 \right),
\]

where \( \tilde{Y}_{i,T+1} \) measures the additional investments, relative to BAU, thanks to the first commitment period. The present-discounted value of this term is \( \frac{a^T}{1 - \delta} \sum_j \tilde{Y}_{i,T+1} \), 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 \( x \) for \( T \) periods at time \( t = 1 \), even if the parties thereafter returned to BAU. The additional utility, relative to BAU, is then:

\[
\sum_{t=1}^{T} \delta^{t-1} \left[ a \sum_{j} (x_{j,t}^{BAU} + x_{j,t}) - \frac{b}{2} \left( x_{i,t}^{BAU} + x_{i,t} - \bar{Y}_{i,t}^{BAU} - Y_{i,t} \right)^2 - \frac{c}{2} \left( \bar{y}_{i,t}^{B} + y_{i,t} \right)^2 - u_{i,t}^{BAU} \right] + a \frac{\delta^T}{1 - \delta} \sum_{j} Y_{j,T+1}
\]

\[
= \sum_{t=1}^{T} \delta^{t-1} \left[ a \sum_{j} x_{j,t} - \frac{b}{2} (x_{i,t} - Y_{i,t})^2 - b \left( x_{i,t}^{BAU} - \bar{Y}_{i,t}^{B} \right) (x_{j,t} - Y_{i,t}) - \frac{c}{2} y_{i,t}^2 - c \bar{y}_{i,t}^{B} y_{i,t} \right] + a \frac{\delta^T}{1 - \delta} \sum_{j} Y_{j,T+1}
\]

\[
= \sum_{t=1}^{T} \delta^{t-1} \left[ a \sum_{j} x_{j,t} - \frac{b}{2} (x_{i,t} - Y_{i,t})^2 - \frac{c}{2} y_{i,t}^2 - b \frac{a}{b} (x_{j,t} - Y_{i,t}) - c \left( \frac{\delta}{1 - \delta} \frac{a}{c} \right) y_{i,t} \right] + a \frac{\delta^T}{1 - \delta} \sum_{j} Y_{j,T+1}
\]

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

where the last equality follows because the three terms with \( Y_{i,\tau} \) in (27) sum to zero for each \( \tau = \{2, ..., T + 1\} \) and because \( Y_{i,1} = 0 \).

When the parties to not play BAU after the first commitment period, then, in order to obtain \( i \)'s full additional payoff relative to BAU, we must add the additional payoff \( \delta^T U_i (x^*) \), where \( U_i (x^*) \) is the equilibrium additional utility relative to BAU, in order to get \( U_i (x) \) in Lemma 1. QED

**Proof of Lemma 2.**

Lemma 1 defines an optimal-control problem with control \( y_{i,t} \). Note that the terminal value for \( Y_{i,T+1} \) is zero because \( U_i (x) \) is measured relative to \( U_i^{BAU} \); 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 \( \lambda \) defines the shadow value of the stock \( Y_{i,t} \), evaluated at time 1, the discrete-time Hamiltonian can be written as:

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

with first-order conditions

\[
y_{i,t} = \arg \max_{y_{i,t}} H_t = \lambda_{i,t+1}/\alpha \delta^{t-1},
\]

adjoint equation

\[
\lambda_{i,t} = \frac{\partial H_t}{Y_{i,t}} = \delta^{t-1} b (x_{i,t} - Y_{i,t}) + \lambda_{i,t+1} \iff
\]

\[
\lambda_{i,t} - \lambda_{i,t+1} = \delta^{t-1} b (x_{i,t} - Y_{i,t}),
\]

37
and terminal condition

\[ \lambda_{i,T+1} = 0 \Leftrightarrow y_{i,T} = 0. \]

Combining the first two conditions and (9), we get the second-order difference equation:

\[
c\delta^{t-2} (Y_{i,t} - Y_{i,t-1}) - c\delta^{t-1} (Y_{i,t+1} - Y_{i,t}) = \delta^{t-1} (x_i - Y_{i,t}) b \Rightarrow
\]

\[-Y_{i,t+1} + \left( \frac{1}{\delta} + 1 + \frac{b}{c} \right) Y_{i,t} - \frac{1}{\delta} Y_{i,t-1} = x_i b/c. \]

A homogeneous solution (when setting the right-hand side equal to zero) is given by \( L^t \), satisfying:

\[
\left[ -L^2 + \left( \frac{1}{\delta} + 1 + \frac{b}{c} \right) L - \frac{1}{\delta} \right] Y_{i,t-1} = 0,
\]

which has the two solutions:

\[
L_1 = \frac{1}{2} \left( \frac{1}{\delta} + 1 + \frac{b}{c} \right) - \frac{1}{2} \sqrt{\left( \frac{1}{\delta} + 1 + \frac{b}{c} \right)^2 - \frac{4}{\delta} \in (0, 1)},
\]

\[
L_2 = \frac{1}{2} \left( \frac{1}{\delta} + 1 + \frac{b}{c} \right) + \frac{1}{2} \sqrt{\left( \frac{1}{\delta} + 1 + \frac{b}{c} \right)^2 - \frac{4}{\delta} > 1}.
\]

When we combine this with the stationary solution

\[ Y_{i,t-1} = x_i, \]

we can write the complete solution as:

\[ Y_{i,t} = \xi_1 L_1^{t-1} + \xi_2 L_2^{t-1} + x_i, \] (28)

where the constants \( \xi_1 \) and \( \xi_2 \) follow from the initial condition \( Y_{i,1} = 0 \), implying \( \xi_1 + \xi_2 = -x_i \), and the terminal condition, \( y_{i,T} = 0 \), implying

\[
y_{i,T} = Y_{i,T+1} - Y_{i,T} = \xi_1 L_1^T \left( 1 - \frac{1}{L_1} \right) + \xi_2 L_2^T \left( 1 - \frac{1}{L_2} \right)
\]

\[
= \xi_1 L_1^T \left( 1 - \frac{1}{L_1} \right) - (\xi_1 + x_i) L_2^T \left( 1 - \frac{1}{L_2} \right) = 0 \Rightarrow
\]

\[
\xi_1 = -\frac{L_2^T \left( 1 - \frac{1}{L_2} \right) x_i}{L_1^T \left( \frac{1}{L_1} - 1 \right) + L_2^T \left( 1 - \frac{1}{L_2} \right)} x_i, \text{ and}
\]

\[
\xi_2 = -\xi_1 - x_i
\]

With the definitions \( C_1 = -\xi x_1 \) and \( C_2 = -\xi x_2 \), (28) can be written as in Lemma 2. QED
Proof of Lemma 3.

By inserting the solutions for $y_{i,t}$ and $Y_{i,t}$ into $U_{i,1}(x)$, defined in Lemma 1, we get:

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

$$= \sum_{t=1}^{T} \delta^{t-1} \left[ a \sum_{j \neq i} x_j - \frac{b}{2} x_i^2 (C_1 L_1^{t-1} + C_2 L_2^{t-1})^2 - \frac{c}{2} x_i (C_1 L_1^{t-1} [1 - L_1] - C_2 L_2^{t-1} [L_2 - 1])^2 \right]$$

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

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

$$\alpha = \sum_{t=1}^{T} \delta^{t-1} a + a \frac{\delta^T}{1-\delta} Y_{j,T+1}$$

$$= a \frac{1 - \delta^T}{1-\delta} + a \frac{\delta^T}{1-\delta} (1 - C_1 L_1^{T-1} - C_2 L_2^{T-1}) = \frac{a}{1-\delta} \left[ 1 - \delta^T (C_1 L_1^{T-1} + C_2 L_2^{T-1}) \right]$$

$$\beta = \sum_{t=1}^{T} \delta^{t-1} \left[ b (C_1 L_1^{t-1} + C_2 L_2^{t-1})^2 + c \left( (C_1 L_1^{t-1} [1 - L_1] - C_2 L_2^{t-1} [L_2 - 1])^2 \right) \right].$$

QED

Proofs of Propositions 1-3.

The proof of Proposition 1 follows from the earlier Lemmata, while Propositions 2 and 3 follow from the reasoning in the text.

Proof of Proposition 4.

If $i$ defects by not contributing at time $t$, then $i$ can still benefit $a \sum_{j \neq i} x_j + \frac{ad}{1-\delta} \sum_{j \neq i} y_{j,t}$, since $j$’s investments will raise $j$’s contribution in the future, even when the parties return to BAU. This benefit is largest at $t = 1$, since $y_{j,t}$ is decreasing in $t \in \{1, ..., T\}$, as noticed already.

When defection is punished by a reversion to BAU for $t' \leq \infty$ periods with probability $p \in [0, 1]$, then compliance (giving payoff $U_i^*$) is better at time $t = 1$ if:

$$a \sum_{j \neq i} x_j + \frac{ad}{1-\delta} \sum_{j \neq i} y_{j,1} + \delta \left( 1 - p + p \delta^{T'} \right) U_i^* \leq U_i^*. $$
When we substitute in for \(y_j, x_j^*,\) and \(U_i^*\), this inequality becomes:

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

\[
a \left[ 1 + \frac{\delta}{1 - \delta} (1 - C_1 L_1 - C_2 L_2) \right] \sum_{j \neq i} x_j^* \leq \left[ 1 - \delta \left( 1 - p + p\delta^t\right) \right] \frac{\alpha^2 (n-1)^2 w}{\beta (1 - \delta^t)} (1 - \frac{w}{2}) \Rightarrow
\]

\[
a \left[ 1 + \frac{\delta}{1 - \delta} (1 - C_1 L_1 - C_2 L_2) \right] (n-1)^2 w \frac{\alpha}{\beta} \leq \left[ 1 - \delta \left( 1 - p + p\delta^t\right) \right] \frac{\alpha^2 (n-1)^2 w}{\beta (1 - \delta^t)} (1 - \frac{w}{2}) \Rightarrow
\]

\[
\frac{a \left( 1 - \delta^t \right)}{\alpha \left[ 1 - \delta \left( 1 - p + p\delta^t\right) \right]} \left[ 1 - \delta (C_1 L_1 + C_2 L_2) \right] \leq 1 - \frac{w}{2} \Rightarrow
\]

\[
w \leq 2 - 2 \frac{1 - \delta (C_1 L_1 + C_2 L_2)}{(1 - \delta) \left[ 1 - \delta \left( 1 - p + p\delta^t\right) \right]} \frac{a \left( 1 - \delta^t \right)}{\alpha},
\]

which equals (18) when \(p = 1\) and \(t' \to \infty\). QED

**Proof of Proposition 5.**

Proposition 5 follows straightforwardly from the equilibrium continuation values, derived above.

**Proof of Proposition 6 (the extensions):**

(i) **Contracts on investments.**

I will first permit the negotiated \(y_t = (y_{1,t}, ..., y_{n,t})\) to be time-dependent, so that \(y = (y_1, ..., y_T)\) is a matrix.

Lemma 1 presents a reformulation of the problem and (when we remove the max-operator) it holds regardless of how the \(x_{i,t}\)'s and the \(y_{i,t}\)'s are decided on. When \(y_{i,t}\) is committed to, but not \(x_{i,t}\), the latter follows straightforwardly from \(i\)'s maximization problem and, just as in BAU,

\[
\bar{x}_{i,t} - \bar{Y}_{i,t} = a/b \Rightarrow x_i = Y_i,t.
\]
The continuation value can thus be written as a function of the vector of investments, $y$:

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

$$U_i(y) - \delta^T U_i(y^*) = \sum_{t=1}^{T} \alpha_t \sum_{j \neq i} y_{j,t} - \frac{\beta_t}{2} y_{i,t}^2,$$

where

$$\alpha_t = \frac{a \delta^{t}}{1-\delta},$$

$$\beta_t = \delta^{t-1} c.$$

$$a \sum_{t=1}^{T} \delta^{t-1} \sum_{j \neq i} (t-1) + a \frac{\delta^T}{1-\delta} \sum_{j \neq i} T y_j = \delta a \frac{1-\delta^T}{(1-\delta)^2},$$

$$\beta = \sum_{t=1}^{T} \delta^{t-1} c = c \frac{1-\delta^T}{1-\delta},$$

so

$$y_i = (n-1) \frac{w}{\beta} = (n-1) w \frac{\delta a/c}{1-\delta}.$$

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

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

where

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

Just as before, we can write $i$’s payoff as in Example E. Consequently, the proofs for the other propositions follow the same steps as above. Proposition 1 gives, for example:

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

**Time-dependent investment levels:** Since $i$’s payoff is separable in the $y_{j,t}$’s, we can apply Corollary 1 for each $y_t$, if we fix the investment levels for the other periods, in order to get:

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

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

In both cases, the choice of $T$ is also irrelevant, since the equilibrium continuation value is (regardless of $T$):

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

$$= \frac{[\delta a (n-1)]^2}{c (1-\delta)^3} w \left( 1 - \frac{w}{2} \right).$$
(ii) Contracts on carbon tax.

I will first permit \( \tau_t = (\tau_{1,t}, \ldots, \tau_{n,t}) \) to be time-dependent, so that \( \tau = (\tau_1, \ldots, \tau_T) \) is a matrix.

With an emission tax equal to \( \tau_{i,t} \), collected by the government in country \( i \), the equilibrium consumption value ensures that the marginal benefit when consuming fossil fuel equals \( \tau_{i,t} \). This implies:

\[
x_{i,t} - Y_{i,t} = \tau_{i,t}/b,
\]

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

\[
U_i(\tau) = \sum_{t=1}^{T} \delta^{t-1} \left[ a \sum_{j \neq i} \tau_{j,t}/b + Y_{j,t} - \frac{\tau_{i,t}^2}{2b} - \frac{c_i}{2} y_{i,t} \right] + a \frac{\delta^T}{1-\delta} \sum_{j \neq i} Y_{j,T+1} + \delta^T U_i(\tau^*)
\]

so, there is no value for \( i \) to invest beyond the BAU-levels, and \( y_{i,t} = 0 \), so:

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

where \( \alpha_t = a\delta^{t-1}/b \), \( \beta_t = \delta^{t-1}/b \).

If the emission tax is time-independent, we can write:

\[
U_i(\tau) - \delta^T U_i(\tau^*) = a \sum_{j \neq i} \tau_j - \frac{\beta}{2} \tau_{i,t}^2,
\]

where

\[
\alpha = a \frac{1 - \delta^T}{b} \frac{1}{1 - \delta},
\]

\[
\beta = \frac{1 - \delta^T}{b} \frac{1}{1 - \delta}.
\]

In this case, Corollary 1 implies:

\[
\tau_{i,t}^* = w (n - 1) \alpha/\beta = w (n - 1) a.
\]

**Time-dependent tax:** Since \( i \)'s payoff is separable in the \( \tau_{j,i} \)'s, we can apply Corollary 1 for each \( \tau_t \), if we fix the emission tax levels for the other periods, in order to get:

\[
\tau_{i,t}^* = w (n - 1) \alpha_t/\beta_t = w (n - 1) a,
\]

which equals \( \tau_{i,t}^* \). Hence, the restriction to time-independent emission tax levels is nonbinding: the equilibrium is the same in both cases.

In both cases, the choice of \( T \) is also irrelevant, since the equilibrium continuation value is (regardless of \( T \)):

\[
U_i(\tau^*) = a \frac{1}{b} \frac{1}{1 - \delta} (n - 1)^2 wa - \frac{11}{2} \frac{1}{b} \frac{1}{1 - \delta} [(n - 1) wa] = a \frac{(n - 1)^2}{b(1 - \delta)} w \left( 1 - \frac{w}{2} \right).
\]
By comparison, a tax gives higher payoff than an investment agreements if:

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

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

(iii) and (iv) Combined. When the parties face both a matrix of emission taxes and a matrix of investment levels, \(i\)'s payoff can be written as:

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

where the first (second) bracket can be recognized 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:

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

Complete contracts. When the parties negotiate the investment levels, the \(\tau_{j,t}\)'s pin down the \(x_{j,t}\)'s, given the \(y_{j,t}\)'s, so negotiating the \(\tau_{j,t}\)'s is then equivalent to negotiating the \(x_{j,t}\)'s: Also when the \(y_{j,t}\)'s and the \(x_{j,t}\)'s 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 \(y_{j,t}\)'s and the \(x_{j,t}\)'s are negotiated, one optimal \(T\) is \(T = \infty\). In this situation, pinning down the \(x_{j,t}\)'s is equivalent to pinning down both the \(y_{j,t}\)'s and the \(x_{j,t}\)'s, because there is no externality when it comes to the \(y_{j,t}\)'s (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:

\[
x_{i,t} = (n-1) \frac{wa}{b} + t(n-1)w \frac{\delta a/c}{1-\delta}.
\]
Given this path, optimal investments, from \( i \)'s point of view, are:

\[
\begin{align*}
\delta cy_{i,t-1} - \delta cy_{i,t} &= \delta b (x_{i,t} - Y_{i,t}) \\
&= \delta b \left( (n-1) wa / b + t (n-1) w \frac{\delta a/c}{1-\delta} - Y_{i,t} \right) \\
&= \delta b \left( (n-1) wa / b + t (n-1) w \frac{\delta a/c}{1-\delta} - t (n-1) w \frac{\delta a/c}{1-\delta} \right) \\
&= \frac{\delta b ((n-1) wa / b)}{c (1-\delta)},
\end{align*}
\]

just as in the optimal contract. So, the combination of negotiating investment levels and emission taxes is 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, and the firms invest to maximize profit, then the outcome coincides with the outcome when all the \( y_{i,t} \)'s and the \( x_{i,t} \)'s are negotiated at the very beginning. (See () and ().)

When only this period's \( x_{i,t} \) are negotiated at the start of period \( t \), then, when applying Corollary 1:

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

Firms invest such as to equalize the marginal investment cost to the present-discounted value of their investment, where the willingness to pay for more \( \bar{y}_{i,t} \) equals \( b (x_{i,t} - \bar{y}_{i,t}) \) at time \( t \). Thus:

\[
c \bar{y}_{i,t} = \sum_{t=1}^{\infty} \delta^t \left( x_{i,t} - \bar{y}_{i,t} \right) = \sum_{t=1}^{\infty} \delta^t b \left( \bar{y}_{i,t}^{BAU} - \bar{y}_{i,t}^{BAU} + x_{i,t} - Y_{i,t} \right)
\]

\[
= \sum_{t=1}^{\infty} \delta^t b \left( \frac{a}{b} + x_{i,t} - Y_{i,t} \right)
\]

\[
= \sum_{t=1}^{\infty} \delta^t b \left( \frac{a}{b} + \frac{a}{b} w (n-1) \right)
\]

\[
= \frac{\delta}{1-\frac{\delta}{c}} \left( \frac{a}{b} + \frac{a}{b} w (n-1) \right).
\]

With \( \bar{y}_{i,t} = \bar{y}_{i,t}^{BAU} + y_{i,t} \) and \( \bar{y}_{i,t}^{BAU} = \frac{\delta}{1-\delta} \frac{a}{c}, \) it follows that

\[
c \bar{y}_{i,t} = \frac{\delta}{1-\delta} aw (n-1),
\]

just as with complete contracts.

**(vii)-(ix)** are trivial and thus omitted.

**(x) Compliance.**

In all the above situations, and also in the basic model if \( c \to \infty \), we have that \( U_i^* \) is independent of \( T \) and it can, when the policy instrument is given by the matrix \( z = (z_1, \ldots, z_K) \), where \( z_k = (z_{1,k}, \ldots, z_{n,k}) \),
for each $k \in \{1, \ldots, K\}$, be written as the following (for some constants $\alpha_k$ and $\beta_k$):

$$U^*_i = \sum_{k \in \{1, \ldots, K\}} \frac{1}{1-\delta} \left[ \alpha'_k \sum_{j \neq i} z_{j,k} - \frac{\beta'_k}{2} z_{j,k}^2 \right]$$

so

$$z_{j,k} = w (n - 1) \frac{\alpha'_k}{\beta'_k},$$

from Corollary 1.

If defection is punished with probability $p$, then the incentive constraint is:

$$\sum_{k \in \{1, \ldots, K\}} \alpha'_k \sum_{j \neq i} z_{j,k} + (1-p) \delta U^*_i \leq U^*_i \Rightarrow$$

$$\sum_{k \in \{1, \ldots, K\}} \alpha'_k w (n - 1)^2 \frac{\alpha'_k}{\beta'_k} \leq \sum_{k \in \{1, \ldots, K\}} \frac{1-(1-p)\delta}{1-\delta} \left[ \alpha'_k w (n - 1)^2 \frac{\alpha'_k}{\beta'_k} - \frac{\beta'_k}{2} \left[ w (n - 1) \frac{\alpha'_k}{\beta'_k} \right]^2 \right] \Rightarrow$$

$$1 \leq \frac{1-(1-p)\delta}{1-\delta} \left[ 1 - \frac{1}{2w} \right] \Rightarrow$$

$$w \leq 2 - 2 \frac{1-\delta}{1-(1-p)\delta} \frac{1-(1-p)\delta}{1-(1-p)\delta} \Rightarrow$$

$$= 2 \delta \left( \frac{1-(1-p)\delta}{1-(1-p)\delta} \right),$$

which simplifies to $w \leq 2\delta$ if $p = 1$.

If, instead, defection is punished for $t'$ periods, then the incentive constraint is equivalent to the one above if simply $1 - p$ is replaced by $\delta''$. QED
References [Preliminary and incomplete; suggestions welcome]


Bodansky and Rajamani: 2017, In book: Evolution and Governance Architecture of the Climate Change Regime, Editors: Detlef Sprinz, Urs Luterbacher,


