Pledge-and-Review Bargaining

25.4.2018. To request the most recent version, please email:

Bård Harstad

bard.harstad@econ.uio.no

Abstract

This paper analyzes a bargaining game that is new to the literature, but that is inspired by real-world international negotiations. With so-called pledge-and-review bargaining, as stipulated in the Paris climate agreement, each party submits an intended nationally determined contribution (INDC), quantifying a cut in its own emission level. Thereafter, the set of pledges must be unanimously ratified. The procedure is repeated periodically, as newly developed technology makes earlier pledges obsolete. I first develop a dynamic model of the pledge-and-review game before deriving four main results: (1) If there is some uncertainty on the set of pledges that is acceptable (for example because of unknown discount factors), each equilibrium pledge is approximated by the *asymmetric Nash Bargaining Solution*. The weights placed on the others' payoffs reflect the underlying uncertainty. The weights vary from pledge to pledge, and the set of equilibrium pledges is inefficient. (2) When this result is embedded in a simple dynamic climate change model, I show that each party contributes too little to the public good, the incentive to develop new technology is weak, and welfare is small. (3) This result is overturned when each party can decide whether or not to participate: The undemanding pledge-and-review process motivates a larger number of parties to participate. This increases aggregate contributions and technology investments, and the parties are better off. (4) When the parties can choose between alternative bargaining games in advance, the (broad but shallow) pledge-and-review game is preferred when there is a larger number of potential participants. The results are consistent with several crucial differences between the climate agreements signed in Kyoto (1997) and Paris (2015) and they rationalize 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. Each party in the negotiation process was first asked to submit an "intended nationally determined contribution" (INDC). After the INDC's had been announced by all parties, the parties were expected to ratify the treaty. The INDC's should specify cuts in the emissions of greenhouse gases being effective from 2020 to 2025 (or to 2030), and every five years the parties shall review and make new pledges for another five-year period.

This negotiation structure 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 compared to the 1990-levels. By comparison, pledge-and-review has been referred to as a "bottom-up" approach since countries will themselves determine how much to cut nationally, without making these cuts conditional on other countries' emission cuts. No wonder, then, that economic theorists question the effectiveness of the pledge-and-review bargaining game.[1]

Interestingly, the Paris agreement differs from the Kyoto Protocol in several other ways, as well: (1) While only 35 countries faced binding emission cuts under Kyoto, the Paris agreement has been signed by nearly every country in the world. (2) While the commitments under the Kyoto Protocol was "legally binding," the Paris Agreement is not. (3) While the Kyoto Protocol was endogenously chosen in the 1990s, the pledge-and-review procedure was favored in the 2010s. At the same time, (4) despite all these differences, the commitment period length was five years for both treaties. Also, both types of agreements share the emphasis on emission cuts rather than specifying national investments in environmentally friendly technology, for example, although the importance of developing such new technology has been emphasized in every recent treaty text.

The purpose of this paper is to propose a framework for studying pledge-and-review bargaining and to use it to shed light on the development from the Kyoto-style agreement to Paris.

The next section describes a bargaining game that is new to the theoretical literature, although it is based on actual (Paris-style) negotiations. With perfect information, the unique and trivial equilibrium of the described bargaining game will coincide with the non-cooperative (or "business-as-usual") outcome, where every party simply contributes so as to maximize its own utility. With sufficiently important shocks on the other parties' willingness to decline and delay ratification, 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 and how shocks are correlated. (The weights also reflect differences in expected discount rates in an intuitive way.) 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. These small weights on others' payoffs makes pledge and review quite different from the (symmetric) Nash Bargaining Solution often used to describe Kyoto-style negotiations.

The subsequent sections investigate the effects of the small weights associated with pledge and review. Section 3 presents a dynamic game in which countries over time make emission cuts as well as invest in new and green technology (or their capacity to make emission cuts in the future). As is consistent with actual negotiations, the parties pledge to make cuts relative to some business-as-usual scenario. The commitments are revised and renegotiated periodically.

Naturally, emission cuts are smaller and (thus) investments in new technology less when the weights are small, for any fixed number of parties. This implies that welfare is also small under pledge-and-review. These conclusions are reversed, however, when the decision to participate in the agreement is endogenized. Since not much is expected from the participating countries (when the weights on others' payoffs are small), it is not that costly to participate and the equilibrium coalition size is larger. The

[1] Also in experiments, pledge-and-review bargaining has lead to disappointing outcomes (Barrett and Dannenberg, 2016).

larger number of parties makes the aggregate cuts more ambitious, the aggregate investments larger, and so is welfare, Section 4 shows.

The outcome under pledge-and-review is clearly quite different than the outcome under the (symmetric) Nash Bargaining Solution, often used to describe the outcome of the Kyoto Protocol. Section 5 compares the outcomes under the two alternative bargaining games when the participation decision is taken into account. 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. In reality, there is a limited number ($\overline{n}$) of potential members and if this upper boundary is binding, then the comparison is less clear. Further, when potential parties are heterogeneous, in that a number ($\underline{n}$) of them will participate regardless of the game, then the "deep and narrow" coalition under the Nash Bargaining Solution can be quite attractive. The result in this section states that pledge-and-review is preferred if and only if $\overline{n}$ is large while $\underline{n}$ is small.

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, $\overline{n}$, has increased. During the same period, eight of the countries that initially signed the Kyoto Protocol have declared that they do not intend to make commitments in the second commitment period of Kyoto: Belarus, Kazakhstan, Ukraine, Japan, New Zealand, Russia, Canada, and USA. This can be interpreted as a smaller $\underline{n}$. Both these developments make pledge-and-review relatively better, according to my theory. Section 5 discusses this in detail, and explain how different groups of countries may have different opinions on exactly when to switch to pledge-and-review.

My model of pledge-and-review bargaining can thus rationalize why many more countries participate in the Paris Agreement than in the Kyoto Agreement (fact 1, above), and why the top-down approach characterizing the Kyoto Protocol was chosen in the 1990s, while pledge-and-review was preferred in the 2010 (fact 3, above). Section 6 also rationalizes why the Kyoto Protocol was legally binding while the Paris Agreement is not (fact 2), and why, despite all these differences, the commitment period length is surprisingly similar between the two systems (fact 4).

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).[2]

The alternating offer bargaining game by Rubinstein (1982) also implements the Nash Bargaining Solution, as shown by Binmore et al. (1986), and asymmetric discount rates rationalizes the *asymmetric* Nash Bargaining Solution (axiomatized by Harsanyi and Selten, 1972; Kalai, 1977; Roth, 1979). Although there can be multiple equilibria with more than two players (Sutton, 1986; Osborne and Rubinstein, 1990), the weights in the asymmetric NBS may then also depend on recognition probabilities (Miyakawa, 2008; Britz et al. 2010; Laruelle and Valenciano, 2008), in addition to the discount rates (Kawamori, 2014).[3]

The next section contributes to this literature by showing that also 'pledge-and-review' bargaining implements the 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.

When applying my model of pledge-and-review bargaining to international treaties (Section 3), I contribute to the literature on whether agreements should be narrow-but-deep or broad-but-shallow (see,

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

[3] 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).

for example, Aldy et al., 2003). 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).[4] This literature predicts that the equilibrium coalition size is very small because of the free-riding incentives. Kolstad and Toman (2005) thus refer to it as a paradox that actual environmental coalitions can often be quite large. There is, however, a well-known tradeoff between treaties that are narrow-but-deep (as in the above-mentioned papers) or broad-but-shallow (Finus and Maus, 2008). My contribution to this literature is to show *why* some bargaining procedures (such as pledge-and-review) lead to shallow (and thus broad) coalitions (Section 4), and to show *when* this procedure is chosen in equilibrium (Section 5). 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 (Section 6). A few extensions are discussed in Section 7 before Section 8 concludes.

[The literature review is preliminary and suggestions are welcome.]

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 several other applications than the climate negotiations motivating the subsequent sections.[5] The main new feature of the game is that each party proposes only one single aspect of the agreement, although payoffs depend on the entire vector of aspects.

There are n parties, each endowed with a payoff function $U_i : \mathbb{R}^n \to \mathbb{R}$, $i \in \{1, ...n\}$. The bargaining game starts when each party i simultaneously proposes its own 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 in the next period, i.e., after some delay, $\Delta > 0$. If everyone accepts, every party i receives the payoff $U_i(\mathbf{x})$ and the game ends.

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). In addition, I will start out by assuming $\partial U_i(\cdot)/\partial x_i < 0$ 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 version of Theorem 1 without these additional assumptions: see the below 'Remark on generality'.

Party i's discount factor between time t and $t + \Delta$ is $\delta_{i,t}^{\Delta}$, but it will be more convenient to refer to the "discount rate" $\rho_{i,t} \equiv \left(1 - \delta_{i,t}^{\Delta}\right)/\Delta$.[6] Thus, i receives $\left(1 - \rho_{i,t}\Delta\right) U_i(\mathbf{x}^*)$ by declining an offer if $\mathbf{x}^*$ can be expected next period. Given $\mathbf{x}^*$, i prefers to accept $\mathbf{x}$ now if:

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

I will restrict attention to stationary subgame-perfect equilibria (SPEs). If information were perfect, it is easy to see that $\mathbf{x}^*$ could be a stationary SPE only if x_i^* were equal to the noncooperative level $x_i^* = \arg\max_{x_i} U_i\left(x_i, \mathbf{x}_{-i}^*\right)$, 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) if just $\rho_{j,t} > 0 \forall j$. Therefore, with the assumptions added above, the "trivial equilibrium" $\mathbf{x}^* = \mathbf{0}$ is 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 discount rates for the next period are not known (to anyone) at the time at which the offers are made. After all, a country's impatience when it

[4] These coalition formation games are rather simple, although the effect of far-sightedness has also been discussed (Ray and Vohra, 2001).

[5] For example, the bargaining game could be appropriate when a number of business partners are negotiating a package deal, and each partner has expertise on and is making the proposal on one single aspect of the package (such as quality, price, delivery time, etc).

[6] If the real discount rate is $\widetilde{\rho}_{i,t}$, the discount factor is $e^{-\widetilde{\rho}_{i,t}\Delta} = \delta_{i,t}^{\Delta}$, so $\rho_{i,t} = \left(1 - e^{-\widetilde{\rho}_{i,t}\Delta}\right)/\Delta$, which approaches $\widetilde{\rho}_{i,t}$ when $\Delta \to 0$. I thus refer to $\rho_{i,t}$ as the discount rate even though the identity holds only in the limit.

comes to ratifying a treaty 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 ρ_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 $\prod_i \left[0, \overline{\theta}_i\right]$, i.i.d. at each time t, and the marginal distribution of $\theta_{i,t}$ is $f_i(\theta_{i,t}) = \int_{\Theta_{-i}} f(\theta_{1,t}, ..., \theta_{n,t})$, where $\Theta_{-i} \equiv \prod_{j \neq i} \left[0, \overline{\theta}_j\right]$. The $\theta_{i,t}$'s are realized and observed by everyone after the offers but before acceptance decisions are made.[7]

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

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

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: $\mathbf{x}^*$ can be sustained as a "nontrivial" stationary SPE if the marginal benefit for i by reducing x_i slightly is outweighed by the risk that at least one party might be sufficiently patient to decline the offer and wait for $\mathbf{x}^*$.

This 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 $\mathbf{x}^*$ is. (The stationary equilibrium $\mathbf{x}^*$ will always be accepted, as is evident from (2) given that $\theta_{i,t} \geq 0$). There can thus be multiple equilibria. To get sharper results, we may want to introduce some small chance that even $\mathbf{x}^*$ will be declined. This will be the consequence if we tremble the support of the discount rate,[8] or if we impose a version of trembling-hand perfection.

Definition of Small Trembles. Assume that when the intended offers are given by $\mathbf{x}$, $\mathbf{x} + \boldsymbol{\epsilon}_t^k$ is realized, where $\boldsymbol{\epsilon}_t^k$ is a vector of n shocks, each i.i.d. over time with mean zero and $\mathrm{E}\left(\boldsymbol{\epsilon}_t^k\right)^2 \to 0$ as $k \to 0$.

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

$$x_i^* \leq \arg\max_{x_i} \prod_{j \in N} \left(U_j\left(x_i, \mathbf{x}_{-i}^*\right)\right)^{w_j^i}, \text{ where } \frac{w_j^i}{w_i^i} = \frac{\rho_i}{\rho_j} f_j(0) \mathrm{E}\left(\theta_{i,t} \mid \theta_{j,t} = 0\right), \forall j \neq i. \tag{3}$$

(ii) If $\mathbf{x}^*(k)$ 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 $\mathbf{x}$ were given by an Asymmetric Nash Bargaining Solution (ANBS), then $\mathbf{x}$ could be described as:

$$x_i^A = \arg\max_{x_i} \prod_{j \in N} \left(U_j\left(x_i, \mathbf{x}_{-i}^A\right)\right)^{w_j} = \arg\max_{x_i} \sum_j w_j \frac{U_j\left(x_i, \mathbf{x}_{-i}^A\right)}{U_j(\mathbf{x}^A)}.$$

For 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 in the pledge-and-review bargaining game, the outcome for 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 every $w_j^i / w_i^i < 1$, it is possible to make every party better off by increasing all the contributions relative to $\mathbf{x}^*$.

[7] This is not unreasonable: (i) Technically, instead of letting $\Delta > 0$ be the delay between rejections and new offers, Δ 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.

[8] In part (ii) of Theorem 1, 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 $\left[\underline{\theta}_j^k, \overline{\theta}_j\right]$ (instead of $\left[0, \overline{\theta}_j\right]$), where $\underline{\theta}_j^k < 0$ and $\underline{\theta}_j^k \uparrow 0 \forall j$ 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.

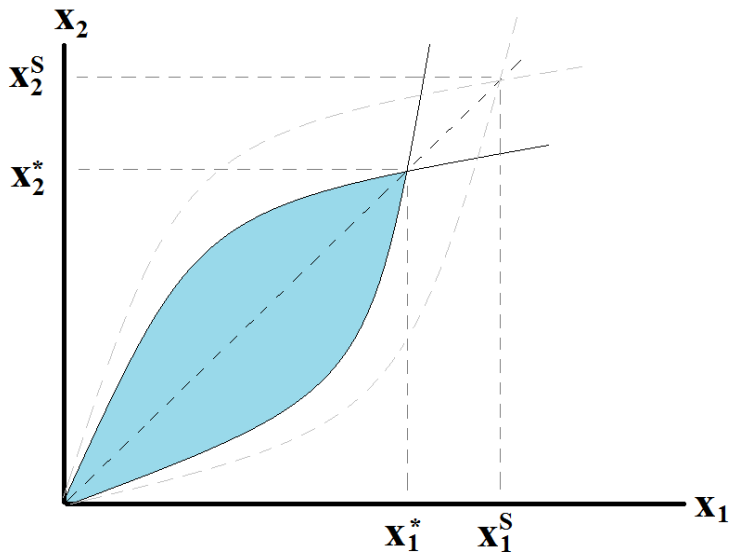

Figure 1: *There are multiple equilibrium contribution levels (unless there are trembles), but they are all smaller than the efficient levels* ($\mathbf{x}^S$).

The theorem also endogenizes the weights and shows how they depend on three thing. 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 $\mathrm{E}(\theta_{i,t} \mid \theta_{j,t} = 0)$, 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.

The theorem has several important consequences.

Corollary 1. *(i) When all parties have the same preferences, the nontrivial* x_i^**'s are simply given by:*

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

where $w = f_j(0) E(\theta_{i,t} \mid \theta_{j,t} = 0) \, \forall i, j$.
(ii) $f_j(0) E(\theta_{i,t} \mid \theta_{j,t} = 0) \leq 1/2 \forall i, j$, if $f_j(\cdot)$ *is single-peaked and symmetric and shocks are not negatively correlated.*[9]

[9] To see this, note that if $f_j(0) > 1/2$, then, when $f_j(\cdot)$ is single-peaked and symmetric around the mean of one, $\int_0^2 f_j(\theta_j)\, d\theta_j > 1$, which is impossible for a pdf $f_j(\cdot)$. If the shocks are i.i.d., then $E(\theta_{i,t} \mid \theta_{j,t} = 0) = 1$. If shocks are positively correlated, then $E(\theta_{i,t} \mid \theta_{j,t} = 0) \leq 1$.

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

These observations are in stark contrast to the symmetric Nash Bargaining Solution, predicting that the x_i's would follow from (4) with $w = 1$.

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:

$$U_i\left(x_i, \mathbf{x}_{-i}^*\right) = \alpha \sum_{j \neq i} x_j - \beta x_i^2/2. \tag{E}$$

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 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.[10] In both cases, the unique nontrivial equilibrium is:

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

Remark on generality. Above, it was assumed that $\partial U_i(\cdot)/\partial x_i < 0$ and $\partial U_j(\cdot)/\partial x_i > 0$, $j \neq i$. Although these assumptions simplify the expression of Theorem 1, a more general version of Theorem 1 is stated and proven in the Appendix and the additional assumptions are not needed for Theorem 2. Further, if the trivial equilibrium, $\mathbf{x}^b$, characterized by $x_i^b = \arg\max_{x_i} U_i\left(x_i; \mathbf{x}_{-i}^b\right) \forall i$, is such that $U_i\left(\mathbf{x}^b\right) > 0$ for some i, then it ceases to exist when we assume the Small Trembles and, therefore, the word "nontrivial" in Theorem 1, part (ii), is, in this case, redundant.

Remark on sufficiency. Condition (3) is necessary for $\mathbf{x}^*$ to be an equilibrium, but it may not be sufficient. Whether the second-order condition for an optimal deviation x_i^i holds globally depends on the pdf f. 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.

3 Implications for Climate Change

To better understand the implications of pledge-and-review, this section embeds the above 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." The 2015 Paris Agreement follows this tradition of letting countries decide on technology themselves. Thus, to be consistent with past and present climate change negotiations, levels of emissions, or emission cuts, are here assumed to be negotiable (and contractible), while technology investments are not. Although this assumption is realistic, it will not be necessary for the main results, as explained in Section ¨7.

For a start, I take as given the set of participants and the length of the commitment period, but both are endogenized in the subsequent sections. Section 5 compares the outcome of pledge-and-review to the outcome under a more traditional (conditional offer) bargaining game, applicable to the Kyoto agreement, in order to determine when pledge-and-review is preferred.

[10] I thank Asher Wolinsky for making this observation.

3.1 A Dynamic Model

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 party j contributes $\widetilde{x}_{i,t}$, the benefit from the sum of contributions is $c\sum_j \widetilde{x}_{j,t}$. This linearity assumption is made for simplicity, but it may also reflects the fact that the marginal benefit from i's contribution at time t is unlikely to change dramatically over (short) periods of time.[11] An additional benefit of this assumption is that we can easily allow for a stock of greenhouse gases that accumulates over time, without changing the analysis, since c can be interpreted as the present discounted-cost of emitting another unit of emission into the atmosphere, when we anticipates that this unit may contribute to climate change for decades.[12]

The second term in the utility function specifies a convex (quadratic) cost of contributing to the public good. It is particularly costly for i to contribute beyond i's capacity level, as measured by the stock $\widetilde{Y}_{i,t}$. This is natural in a climate change model, where a country can consume energy from both fossil fuels ($g_{i,t}$) as well as renewables ($\widetilde{Y}_{i,t}$). If the total consumption of energy is less than i's bliss point, $\widehat{Y}_i$, then i experiences a disutility $\frac{b}{2}\left(\widehat{Y}_i - \left[g_{i,t} + \widetilde{Y}_{i,t}\right]\right)^2$. This disutility can be written as $\frac{b}{2}\left(\widetilde{x}_{i,t} - \widetilde{Y}_{i,t}\right)^2$, if we let $\widetilde{x}_{i,t} \equiv \widehat{Y}_i - g_i$ measure i's emission cut relative to $\widehat{Y}_i$.

Each party can over time invest in and add to its stock, $\widetilde{Y}_{i,t}$. The investment cost is assumed to be convex and quadratic and characterized by parameters K_i and k. The investment cost is the third term in the per-period utility function:

$$\begin{aligned} \widetilde{u}_{i,t} &= c\sum_j \widetilde{x}_{i,t} - \frac{b}{2}\left(\widetilde{x}_{i,t} - \widetilde{Y}_{i,t}\right)^2 - \frac{k}{2}\left(K_i + \widetilde{y}_{i,t}\right)^2, \text{ where} \qquad (5) \\ \widetilde{Y}_{i,t+1} &= \widetilde{Y}_{i,t} + \widetilde{y}_{i,t}. \end{aligned}$$

The parties can have heterogeneous bliss points for consumption, initial technology levels ($\widetilde{Y}_{i,0}$), and investments costs (K_i). For simplicity, the parties are assumed to be identical in other respects and they plan by applying the same (expected) discount factor, δ.[13]

As a benchmark, consider the noncooperative MPE without bargaining, referred to as the "business as usual" (BAU) equilibrium. The pollution stock is not payoff-relevant, and thus every $\widetilde{x}_{i,t}$ will satisfy the first-order condition $b\left(\widetilde{x}_{i,t} - \widetilde{Y}_{i,t}\right) = c$. This, in turn, implies that each investment unit will lead to another unit of $\widetilde{x}_{i,t}$ for every future period, without influencing $\widetilde{x}_{i,t} - \widetilde{Y}_{i,t}$. Hence, the stock $\widetilde{Y}_{i,t}$ is also payoff-irrelevant and every $\widetilde{y}_{i,t}$ must satisfy the first-order condition $k\left(K_i + \widetilde{y}_{i,t}\right) = c\delta/\left(1-\delta\right)$, assuming $c\delta/\left(1-\delta\right)k - K_i > 0$. (I.e., I assume that the constant K_i is so small that investments are positive. Parameter K_i will play no role in the analysis and we may just as well set it equal to zero.) The second-order conditions hold trivially, so the BAU is:

$$\widetilde{x}_{i,t}^{BAU} = \widetilde{Y}_{i,t} + \frac{c}{b} \text{ and } \widetilde{y}_{i,t}^{BAU} = c\delta/\left(1-\delta\right)k - K_i.$$

Now, consider the possibility that the parties will contribute *more* than the BAU-levels. In particular, suppose i agrees at time zero to contribute $x_i \geq 0$ units, beyond i's BAU level, for each of the next T periods. These contributions may motivate i to invest $y_{i,t}$ units, in addition to the business-as-usual level,

[11] Golosov et al. (2014) estimate that the climate damage function is approximately linear.

[12] To see this, suppose pary i emits $g_{i,t}$ and the pollution stock is $G_t = q_G G_{t-1} + \sum_j g_{j,t}$, depreciating at rate $1 - q_G \in [0,1]$. If parameter $C > 0$ measures each party's per-period marginal environmental harm from the stock G_t, then the present-discounted cost of another unit of emission is $c \equiv C/\left(1 - q_G\delta\right)$. Consequently, this c also measures the marginal present-discounted benefit from a cut in the emission level.

[13] Thus, i seeks to maximize $\sum_{t=0}^{\infty}\delta^t\widetilde{u}_{i,t}$ at time 0. Even if a party's impatience was allowed to be stochastic and uncertain during the bargaining process (this was assumed 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 be related to policy makers' need to give attention to other urgent policy or economic issues. These shocks cannot be predicted in advance and for long-term technology investments, one will apply the expected discount factors (since the utility is linear in the discount factor, and there is no risk aversion w.r.t. them).

at each time t. Total contributions and investments are then:

$$\widetilde{x}_{i,t} = \widetilde{x}_{i,t}^{BAU} + x_i \text{ and } \widetilde{y}_{i,t} = \widetilde{y}_{i,t}^{BAU} + y_{i,t}.$$

Note that the additional contribution, x_i, is assumed to be constant for every time within the same commitment period. We could alternatively have assumed that the contributions were time-dependent, or that the pledge pinned down an absolute level for the contribution that would be binding in every period (as in the previous version of this paper). Section 7 argues that these alternative assumptions on the nature of the pledge would *not* alter the main results.

The timing of the game is the following. First, the parties negotiate the contributions (relative to BAU), x_i, pinning down the contribution levels for the next T periods $\{0, 1, ..., T-1\}$. 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.[14]

I will now derive 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, Theorem 1 can be applied to characterize the bargaining outcome for the x_i's and thus the implications of pledge-and-review for equilibrium investment levels. I start by treating as exogenous parameters n, w, and T, since these might be determined by forces outside of this model, but all these parameters are endogenized in Sections 4-7.

Other extensions: Section 7 also discusses how the model easily can be extended to situations in which investments or contractible, or made by firms rather than governments, and if the negotiated contribution levels are allowed to be functions of time, $\{x_{i,t}\}$.

3.2 The Optimal Control Problem

Since the BAU levels are constant, we can use them as a benchmark when measuring the impact of the additional contributions. With x_i, i's payoff at time t is:

$$\begin{aligned}
\widetilde{u}_{i,t}^{x} &= c\sum_j \left(\widetilde{x}_{j,t}^{BAU} + x_j\right) - \frac{b}{2}\left(\widetilde{x}_{j,t}^{BAU} + x_i - \widetilde{Y}_{i,t}^{BAU} - Y_{i,t}\right)^2 - \frac{k}{2}\left(K_i + \widetilde{y}_{j,t}^{BAU} + y_{i,t}\right)^2 \Rightarrow \\
u_{i,t} \equiv \widetilde{u}_{i,t}^{x} - \widetilde{u}_{i,t}^{bau} &= c\sum_j x_j - \frac{b}{2}\left(x_i - Y_{i,t}\right)^2 - \frac{k}{2}y_{i,t}^2 + \left[Y_{i,t}c - y_{i,t}c\delta/\left(1-\delta\right)\right], \text{ where} \\
Y_{i,t+1} &= Y_{i,t} + y_{i,t} \text{ and } Y_{i,0} = 0.
\end{aligned}$$

Note that the heterogenous parameters in (5) drops out when the utility is measured relative to BAU.

Once the contributions (the x_i's) have been negotiated, each party faces an optimal control theory problem when it comes to the investment levels for the next T periods.

3.3 Equilibrium Investments

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

[14] The pledges in the Paris Agreement will be renewed every five years. The motivaiton is that "with ever better technology and with a much greater flow of financing to developing countries, the ambition of these contributions, which will be revisited after "stocktakes" every five years, will quickly grow." https://www.economist.com/news/international/21683990-paris-agreement-climate-change-talks

Lemma 1. *In equilibrium, the stock $Y_{i,t}$ and the investment $y_{i,t}$ are both linear in x_i:*

$$\begin{aligned}
Y_{i,t} &= x_i\left(1 - C_1 L_1^t - C_2 L_2^t\right), \text{ and, therefore,} \\
y_{i,t} &= x_i\left[C_1 L_1^t (1 - L_1) - C_2 L_2^t (L_2 - 1)\right], \text{ where} \\
L_1 &\equiv \frac{1 + 1/d + b/k}{2} - \sqrt{\left(\frac{1 + 1/d + b/k}{2}\right)^2 - \frac{1}{d}} \in (0,1) \\
L_2 &\equiv \frac{1 + 1/d + b/k}{2} + \sqrt{\left(\frac{1 + 1/d + b/k}{2}\right)^2 - \frac{1}{d}} > 1, \\
C_1 &\equiv \frac{L_2^{T-1}(L_2 - 1)}{L_2^{T-1}(L_2 - 1) + L_1^{T-1}(1 - L_1)} \in (0,1), \\
C_2 &\equiv \frac{L_1^{T-1}(1 - L_1)}{L_2^{T-1}(L_2 - 1) + L_1^{T-1}(1 - L_1)} = 1 - C_1 \in (0,1).
\end{aligned}$$

Naturally, if i is committed to contribute a lot, in that x_i is large, then i invests more. In a climate change setting, a commitment to large emission cuts motivates the countries to invest in renewables.

It is also easy to check that $y_{i,t}$ decreases over time and reaches zero in the last period. Consequently, if $T = 1$, then $y_{i,0} = Y_{i,0} = 0$. 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 $(y_{i,T-1})$ 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 (therefore) investment levels will be linear functions of x_i. Thus, 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, as proven in the Appendix:

Lemma 2. *Since every $y_{i,t}$ is linear in x_i, i's continuation value, relative to BAU, can be written as in Example E:*

$$\begin{aligned}
U_i(\mathbf{x}) = \sum_{t=0}^{\infty} \delta^t u_{i,t} &= \alpha \sum_{j \neq i} x_j - \frac{\beta}{2} x_i^2, \qquad \text{(E)} \\
&\text{where } \alpha \text{ and } \beta \text{ are defined as} \\
\alpha &\equiv \frac{c}{1 - \delta}\left[1 + \delta^T\left(1 - C_1 L_1^T - C_2 L_2^T\right)\right], \\
\beta &\equiv \sum_{t=0}^{T-1} \delta^t \left[\frac{b}{2}\left(C_1 L_1^t + C_2 L_2^t\right)^2 + \frac{k}{2}\left(C_1 L_1^t [1 - L_1] - C_2 L_2^t [L_2 - 1]\right)^2\right].
\end{aligned}$$

Thus, the continuation value $U_i(\mathbf{x})$ simplifies to Example E, introduced in Section 2, even though α and β are relatively complicated functions of b, c, k, δ, and T.

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

$$x_i^* = \arg\max_{x_i} U_i(\mathbf{x}) + w \sum_{j \neq i} U_j(\mathbf{x}) = w(n-1)\alpha/\beta. \qquad (6)$$

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 (6) and (E), we

can see that U_i is increases in w for every $w \leq 1$:

$$U_i = \alpha (n-1) \left[w (n-1) \frac{\alpha}{\beta} \right] - \frac{\beta}{2} \left[w (n-1) \frac{\alpha}{\beta} \right]^2 = \frac{\alpha^2}{\beta} (n-1)^2 w \left(1 - \frac{w}{2} \right). \tag{7}$$

Proposition 1. *A smaller w reduces contributions, investments, and payoffs.*

4 Participation

While the number of participants so far has been taken as exogenous, this section endogenizes the coalition size and studies how it depends on the pledge-and-review bargaining procedure.

A standard 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, $i \in \{1, ..., \overline{n}\}$, 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, while the free-riders will simply follow their business-as-usual strategy. It is most natural and standard to focus on pure-strategy equilibria at the participation stage, and doing so pins down the equilibrium participation number, n.

Since the parties' payoffs in the climate change model can be summarized as (E), according to Lemma 2, we will henceforth restrict attention to these payoffs. Thus, the larger is the equilibrium n, the larger is each participant's contribution level, while every free-rider sets $x_i = 0$.

I start by ignoring the constraint $n \leq \overline{n}$; but that constraint is extensively discussion in the next section.

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, and 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 w of) the utility of one additional coalition member.

The payoff for each of the n participants is given by (7). If one of these parties instead free rides, the party's payoff will be $\alpha (n-1) w (n-2) \alpha/\beta$, since each of the other parties will now contribute $w (n-2) \alpha/\beta$. By comparison, participation is beneficial if:

$$\begin{aligned} \frac{\alpha^2}{\beta} (n-1)^2 w (1 - w/2) &\geq \alpha (n-1) w (n-2) \alpha/\beta \Rightarrow \\ (n-1) w &\leq 2. \end{aligned} \tag{8}$$

Thus, for n to be an equilibrium coalition size, (8) must hold for the equilibrium n, but it must fail for any larger n (since, otherwise, additional members would like to participate).

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

$$n = \left\lfloor 1 + \frac{2}{w} \right\rfloor.$$

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

Importantly, n decreases in w. If $w = 1$, as when applying NBS, $n = 3$, according to Proposition 2: this is the well-know result in the literature focusing on linear-quadratic utility functions, mentioned above. However, when w is smaller, as in the pledge-and-review bargaining game, then a coalition member is not expected to contribute a lot. This reduces the cost of participation, and participation is attractive for a larger set of n's. The size n cannot be too large, however, since then individual contributions would be so large that free riding would be preferable.

Since the number of participants must be an integer, n is a step-function that decreases in w. The contribution level, as given by, will thus increase in w when w increases so little that n stays unchanged. For a sufficiently large increase in w, n drops, and so does x_i.

When comparing pledge-and-review to NBS, we are interested in large rather than small differences in w. Thus, it is not unreasonable to ignore the fact that n must be an integer and apply the approximation $\lfloor z \rfloor \approx z$. This implies:

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

With this approximation, the product $(n(w) - 1)w$ stays constant if w changes. To understand why, 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 participating remains unchanged. 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 a the cost (and therefore the indifference condition) are unchanged when w decreases if just n simultaneously increases so much that the product $(n-1)w$ is unchanged.

4.2 Implications for Climate Change

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 larger, overall welfare will be larger when w is small. This is evident when the endogenization of n, as described by (9), is combined with (7). This gives:

$$U_i = 4\frac{\alpha^2}{\beta}\left(\frac{1}{w} - \frac{1}{2}\right). \tag{10}$$

Proposition 3. *With endogenous n, and (9), Proposition 1 is reversed: A smaller w increases aggregate contributions, investments, and welfare.*

5 When to Choose Pledge-and-Review

Pledge-and-review bargaining is associated with a relatively small w, according to Theorem 1 and Corollary 1. A small w reduces the participants' payoffs when n is taken as exogenous but not when n is endogenous, according to Propositions 1 and 3. Thus, a treaty that is "shallow but broad" (in that w is small but n large) is preferable to a treaty that is "deep but narrow" (in that w is large but n small).

In reality, there are several reasons for why n does not always increase according to $n(w)$, defined by (9), when w declines. First, the world consists of a finite number of countries, $\overline{n}$. When w is so small that everyone participates, $\overline{n} \leq n(w)$, then a further reduction in w is harmful for everyone, just as described by Proposition 1. According to this argument, overall welfare (and aggregate contributions and investments) is single-peaked in w and maximized when everyone is just willing to participate: $n(w) = \overline{n}$.

Second, countries are more heterogeneous in reality than the model above has permitted. If the willingness to participate had varied across countries, the countries that would benefit most from participating are the first ones to participate if w is reduced. Countries that benefit less from participating would not be indifferent and a significant reduction in w would therefore be necessary for them to be willing to participate. In this case, $(n-1)w$ is likely to increase in w when countries are heterogeneous.

There are several ways of capturing this reasoning in the model. Analytically, the simplest way is to assume that some parties 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. The European Union member countries, for example, cannot easily opt out of an environmental agreement unilaterally.

Let $\underline{n} \leq \overline{n}$ measure the number of committed parties. Thus, exactly $\underline{n}$ will participate when $\underline{n} \geq n(w)$, where $n(w)$ is defined above.

For either of these two reasons, a bargaining game associated with a small w is not necessarily superior to a bargaining game associated with a large one. To understand when pledge-and-review is preferable, this section compares the participants' payoffs when $w = \underline{w}$ and when $w = \overline{w} > \underline{w}$.

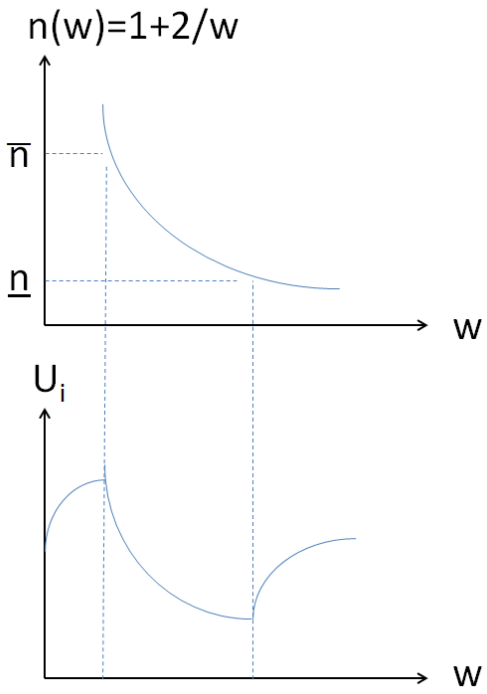

Figure 2: *Participation and participants' payoffs are strictly decreasing in w only when $n(w) \in (\underline{n}, \overline{n})$.*

5.1 The Preferred Bargaining Game

From (9) it follows that $n(\underline{w}) > n(\overline{w})$. If either $n(\overline{w}) > \overline{n}$ or $n(\underline{w}) < \underline{n}$, the number of participants is the same for the two games and thus Proposition 1 shows that $\overline{w}$ is preferable. If instead $\underline{n} < n(\overline{w}) < n(\underline{w}) < \overline{n}$, the constraints are not binding and Proposition 3 shows that $\underline{w}$ is preferable. There is an interesting trade-off only in three situations.

(a) $\overline{n}$ binds: Suppose $\underline{n} < n(\overline{w}) < \overline{n} < n(\underline{w})$. In this case, $\underline{w}$ leads to full participation while participation under $\overline{w}$ is given by $n(\overline{w})$. Thus, a sufficiently large $\overline{n}$ ensures that a participant's payoff is largest under $\underline{w}$. From (7):

$$\begin{aligned} \frac{\alpha^2}{\beta}(\overline{n}-1)^2\underline{w}^2\left(\frac{1}{\underline{w}}-\frac{1}{2}\right) &> \frac{\alpha^2}{\beta}(n(\overline{w})-1)^2\overline{w}^2\left(\frac{1}{\overline{w}}-\frac{1}{2}\right) \Rightarrow \\ \frac{\overline{n}-1}{n(\overline{w})-1} &> \Omega, \text{ where} \\ \Omega &\equiv \sqrt{\frac{\overline{w}(1-\overline{w}/2)}{\underline{w}(1-\underline{w}/2)}} \in \left(1, \frac{\overline{w}}{\underline{w}}\right). \end{aligned}$$

(b) $\underline{n}$ binds: Suppose $n(\overline{w}) < \underline{n} < n(\underline{w}) < \overline{n}$. In this case, only committed parties participates under $\overline{w}$, while participation under $\underline{w}$ is given by $n(\underline{w})$. Thus, a sufficiently small $\underline{n}$ ensures that a participant's payoff is largest under $\underline{w}$. From (7):

$$\begin{aligned} \frac{\alpha^2}{\beta}(n(\underline{w})-1)^2\underline{w}^2\left(\frac{1}{\underline{w}}-\frac{1}{2}\right) &> \frac{\alpha^2}{\beta}(\underline{n}-1)^2\overline{w}^2\left(\frac{1}{\overline{w}}-\frac{1}{2}\right) \Rightarrow \\ \frac{n(\underline{w})-1}{\underline{n}-1} &> \Omega. \end{aligned}$$

(c) Both $\underline{n}$ and $\overline{n}$ bind. Suppose $n(\overline{w}) < \underline{n} < \overline{n} < n(\underline{w})$. In this case, there is full participation under $\underline{w}$, but only the committed parties participate under $\overline{w}$. In this situation, $\underline{w}$ is preferred when $\overline{n}$ is large

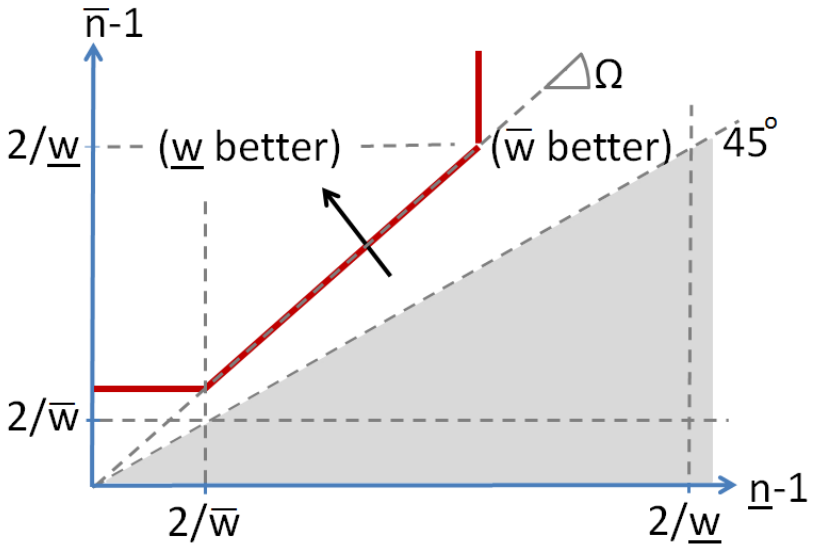

Figure 3: *Participants prefer to switch to pledge-and-review ($\underline{w}$) above the solid line.*

while $\underline{n}$ is small. From (7):

$$\begin{aligned} \frac{\alpha^2}{\beta}(\overline{n}-1)^2\underline{w}^2\left(\frac{1}{\underline{w}}-\frac{1}{2}\right) &> \frac{\alpha^2}{\beta}(\underline{n}-1)^2\overline{w}^2\left(\frac{1}{\overline{w}}-\frac{1}{2}\right) \Rightarrow \\ \frac{\overline{n}-1}{\underline{n}-1} &> \Omega. \end{aligned}$$

Since $n(w)-1=2/w$, the three cases can be combined as follows.

Proposition 4. *Participants prefer switching to pledge-and-review (i.e., from $w=\overline{w}$ to $w=\underline{w}<\overline{w}$) if $\overline{n}$ is large while $\underline{n}$ is small. The exact condition is:*

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

This condition is better illustrated in a figure. If there is a larger number of potential parties, or if fewer countries are committed to participate even when w is large, then we may move in the direction of the arrow in Figure 3, so that the "shallow" agreement becomes preferred even though the "deep" agreement was preferred for a smaller number of potential parties, or for a larger number of committed parties.

5.2 From Kyoto to Paris

One may argue that both these developments (i.e., a larger $\overline{n}$ and a smaller $\underline{n}$) are in line with changes in world politics the last few decades. Today we have a larger number of countries that are emerging economies, although they up to recently 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, $\overline{n}$, has increased.

During the same period, eight 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 periods: Belarus, Kazakhstan, Ukraine, Japan, New Zealand, Russia, Canada, and USA. This reluctance may be interpreted as a reduction in $\underline{n}$.

For either (or both) reason, the switch to pledge-and-review is thus consistent with the theory of this paper.

To get a sense of when pledge-and-review is preferred, suppose the Kyoto Protocol is described by NBS with $\overline{w} = 1$ and $n = 35$. When the Paris Agreement attracts 195 participants, it is preferred according to Proposition 4 when:

$$\frac{195-1}{35-1} > \sqrt{\frac{1\left(1-1/2\right)}{\underline{w}\left(1-\underline{w}/2\right)}} \Leftrightarrow \underline{w} > 0,015.$$

Thus, the participants prefer to switch to pledge-and-review even if this bargaining procedure motivates the parties to take into account only 1.5 percent of the other parties' utilities.

It is straightforward to show that the committed countries prefer to switch to a shallow agreement for a larger set of parameters than do the uncommitted countries. To see the intuition for this, note that the participants benefit in two ways from a larger n. First, a larger n might lead to larger aggregate contributions. This benefit is clearly present if $\underline{n}$ and $\overline{n}$ are not binding $n\left(w\right)$, so that $n\left(w\right) \in (\underline{n}, \overline{n})$, but a larger n (motivated by a smaller w) can also lead to *less* aggregate contributions, when one of the constraints bind. 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 $w = \overline{w}$. Thus, the original set of participants may switch to pledge-and-review too early, i.e., at a stage when this switch reduces 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.

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 Section 4, 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.

When providing public goods internationally, there is no third party or "world government" that can force participants to contribute as promised. Therefore, many scholars argue that a treaty must be self-enforcing, i.e., that it must be in the interest of each participant to contribute as pledged, even when the possibility to free ride arises (see, for example, Barrett (1994), Dutta and Radner (2004), or Harstad et al. (2018)).

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. Thereafter, however, there is a risk that future cooperation will suffer. To investigate the extent to which this risk can motivate compliance, suppose we revert to the noncooperative MPE – the BAU – as soon as one party has defected by contributing less than pledged. Note that this is the worst possible threat if the parties cannot observe the identity of the party that defected. (If everyone knew the identity of the party that defected, then minmax strategies could be worse than BAU for the defecting party, and a treaty would be self-enforcing for a larger set of circumstances than those henceforth derived.)

Consider first a simple repeated game in which n parties contribute x_i in every period and continuation payoffs are given by Example E and (7). If a participant defects by not contributing, this party receives the (normalized) payoff $\left(1-\delta\right)\alpha\sum_{j\neq i}x_j = \left(1-\delta\right)\left(n-1\right)^2 w\alpha^2/\beta$ one period, and nothing thereafter. Defection is thus not attractive if:

$$\begin{aligned} \frac{\alpha^2}{\beta}\left(n-1\right)^2 w\left(1-w/2\right) &> \left(1-\delta\right)\left(n-1\right)^2 w\alpha^2/\beta \Rightarrow \\ w &< 2\delta. \end{aligned}$$

This condition holds if the discount factor is large and the treaty rather shallow.

The dynamic climate change game in Section 3 is somewhat different, since the investments influence the 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, even if the parties are then reverting to the noncooperative MPE. The (normalized) payoff when defecting at time t is thus as expressed on the right-hand side in the following compliance constraint:

$$\begin{aligned} \frac{\alpha^2}{\beta}(n-1)^2 w(1-w/2) &> (1-\delta)c\left[\sum_{j\neq i} x_{j,t} + \frac{\delta}{1-\delta}\sum_{j\neq i} y_{j,t}\right] \Rightarrow \\ w &< 2 - 2(1-\delta)\frac{(1-\delta) + \delta\left[C_1 L_1^t (1-L_1) - C_2 L_2^t (L_2-1)\right]}{1+\delta^T\left(1 - C_1 L_1^T - C_2 L_2^T\right)}. \end{aligned}$$

The implication follows when we substitute in for the equilibrium $y_{i,t}$, $x_{i,t}$, and α. This condition is hardest to satisfy when t is small, since investments are largest at the beginning of each commitment period.

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

Proposition 5. *The bargaining outcome is more likely to be self-enforcing if w is small. This claim holds whether or not participation is endogenous, and whether or not the parties invest over time.*

If 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. The fact that the Kyoto Protocol was "legally binding" is likely to raise the political cost if one does not comply later on. The contributions following the Paris Agreement, in contrast, is not legally binding. This difference between the two treaties is consistent with the proposition above: Since the Paris Agreement applies pledge-and-review bargaining, where w is smaller, it is more likely that (**??**) holds for this agreement without making it legally binding.

7 The Commitment Period Length — And Extensions

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 to believe that the second commitment period's length of *eight* years was agreed to only because a global treaty was under planning for 2020 (this became the Paris Agreement).

This similarity may be surprising because the period length is not irrelevant. In particular, 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 1. The intuition for this comparative static is the standard hold-up problem: if the next bargaining stage (at T) 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 (just as in Harstad, 2016). On the other hand, after investments have been made, it would be better for all the parties to start the pledge-and-review bargaining game soon again, to take advantage of the newly developed technology. When T is committed to in advance, the optimal T trades off the positive effect on investments and the benefit that newly developed technology can strengthen the commitments sooner when T is small.

Despite the fact that there is a trade-off when it comes to deciding on the optimal T, the optimal T is independent of w and n in the model above. Proposition 2 has already shown that the equilibrium n is independent of T^*, and Proposition 4 showed that the choice of w is independent of T^*. When n is exogenous, then a party's payoff is given by (7), where both α and β are complicated functions of T.

When n is instead endogenous, a party's payoff is given by (10). The optimal contract length is the same in both cases:

$$T^* = \arg\max_T \frac{\alpha^2}{\beta}(n-1)^2 w\left(1-\frac{w}{2}\right) = \arg\max_T 4\frac{\alpha^2}{\beta}\left(\frac{1}{w}-\frac{1}{2}\right) = \arg\max_T \frac{\alpha^2}{\beta}.$$

Proposition 6. *The optimal commitment period length, T^*, is independent of n and w, and of whether n and w are endogenous or exogenous.*

This result shows that there may be no reason to change the optimal commitment period length, despite the many other differences between two treaties. This is consistent with the comparison between the Kyoto Protocol and the Paris Agreement.

Needless to say, the choice of T 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. The results above have therefore been derived for any fixed T, and they hold regardless of the choice of T. The rest of this section discusses how the choice of T will be modified if we change some of the assumptions above, even though the propositions above continue to hold.

Firms investing. If firms, rather than countries, decide on the investment levels, and if countries are unable to regulate the firms' investments, then all results continue to hold except that the optimal commitment period length is reduced to one. The reason for this is that firms would not invest less when the next bargaining game is close in time, since it would be countries and not firms that would be subject to the hold-up problem described above. This result would hold regardless of n and w, and of whether n and w are endogenous or exogenous. Thus, Proposition 6 will continue to hold, and so will Propositions 1-5. Of course, if each government can subsidize/tax the firms' investments, then the government can implement its preferred choice of investment, as described in Section 3, and then even the exact equations above stay unchanged.

Committing to a path of contributions. Above, it has been assumed that the parties pledge contributions (relative to BAU) that stay constant until time T, when new pledges will be negotiated. If, instead, parties negotiated an arbitrary sequence of additional contribution levels, $\{x_{i,t}\}$, then there would be no reason to update the pledges frequently. In this case, the optimal T would be infinite, since the long T is beneficial for avoiding the hold-up problem. Again, this result holds regardless of n and w, and of whether n and w are endogenous or exogenous. Thus, Propositions 1-6 continue to hold.

Contractible investments. If the parties can contract on contribution levels as well as investment levels, then there is no hold-up problem and thus no reason to have a long T. In this case, the optimal T equals one. If investments are contractible and the parties can also negotiate a path of contributions, $\{x_{i,t}\}$, then the choice of T will be irrelevant, it can be shown.

Although these three extensions leave the results above unchanged, it is certainly possible to think off other forces that would make the choice of T depend on w or n. In Harstad (2016), I introduce shocks on the marginal environmental harm, and these shocks accumulates 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), the countries that have decided to participate decides on T after observing n. Then, they may prefer a small T if n is small, since the small T facilitates the admission of new participants sooner. While these and alternative extensions may predict that T should be a function of the bargaining procedure, they are not necessary to rationalize the above-mentioned differences between the Kyoto Protocol and the Paris Agreement.

8 Concluding Remarks

The novelty of pledge-and-review bargaining is that each party proposes how much to contribute individually – unconditional on what other parties pledge – before the vector of pledges must be agreed to by all the parties. 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. The two treaties are also different in other respects: (1) Many

more countries contribute to the Paris Agreement than to the Kyoto Protocol, (2) while the commitments under the Kyoto Protocol was legally binding, the Paris Agreement is not, (3) while the Kyoto Protocol was endogenously chosen in the 1990s, the pledge-and-review procedure was favored in the 2010s, but (4) despite all these differences, the commitment period length is similar for both treaties.

This paper provides a model of pledge-and-review bargaining game. 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. It was shown that each party's contribution level is as described by an asymmetric Nash Bargaining Solution. Since the weights vary from pledge to pledge, the collection of pledges is not Pareto optimal. Abatement levels will be small and so will investments in new and "green" technology. However, these results are reversed when the coalition size is endogenous: then, the shallowness of pledge-and-review makes it affordable for a larger number of countries to participate, and this effect dominates the fact that each party places a smaller weight on the utility of the others. The model can thus motivate the use of pledge-and-review. More importantly, the predictions of the model is consistent with the other differences between the Kyoto Protocol and the Paris Agreement, as summarized by (1)-(4).

9 Appendix [Preliminary]

Proof of Theorem 1.

As advertised in Section 2, the following version of Theorem 1(i) 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(i)^G. *If x^* is a nontrivial stationary SPE in which $U_i(\mathbf{x}^*) > 0 \forall i$, then, for every $i \in N$, we have:*

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

$$-\frac{\partial U_i(\mathbf{x}^*)}{\partial x_i} \leq \sum_{j \setminus i} \max\left\{0, \frac{\partial U_j(\mathbf{x}^*)/\partial x_i}{\rho_j \Delta U_j(\mathbf{x}^*)}\right\} f_j(0)\, \mathrm{E}(\theta_{i,t} \mid \theta_{j,t} = 0)\, \rho_i \Delta U_i(\mathbf{x}^*); \tag{11}$$

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

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

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 (11).

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

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

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

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

$$\begin{aligned} \left(1 - p_{-j}\left(\mathbf{x}^i\right)\right) U_j\left(\mathbf{x}^i\right) + p_{-j}\left(\mathbf{x}^i\right)\left(1 - \rho_{j,t}\Delta\right) U_j(\mathbf{x}^*) \quad &< \quad \left(1 - \rho_{j,t}\Delta\right) U_j(\mathbf{x}^*) \Rightarrow \\ \theta_{j,t} < \widetilde{\theta}_j\left(\mathbf{x}^i\right) \equiv \max\left\{0, \frac{U_j(\mathbf{x}^*) - U_j\left(\mathbf{x}^i\right)}{\rho_j \Delta U_j(\mathbf{x}^*)}\right\}. & \end{aligned}$$

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

$$\begin{aligned} 1 - p\left(\mathbf{x}^i\right) &= G\left(\widetilde{\theta}_1\left(\mathbf{x}^i\right), ..., \widetilde{\theta}_{i-1}\left(\mathbf{x}^i\right), \widetilde{\theta}_{i+1}\left(\mathbf{x}^i\right), ...\widetilde{\theta}_n\left(\mathbf{x}^i\right)\right) \\ &\equiv \int_0^{\overline{\theta}_i} \left[\int_{\widetilde{\theta}_1(\mathbf{x}^i)}^{\overline{\theta}_1} ... \int_{\widetilde{\theta}_{i-1}(\mathbf{x}^i)}^{\overline{\theta}_{i-1}} \int_{\widetilde{\theta}_{i+1}(\mathbf{x}^i)}^{\overline{\theta}_{i+1}} ... \int_{\widetilde{\theta}_n(\mathbf{x}^i)}^{\overline{\theta}_n} f(\boldsymbol{\theta}_t)\, d\boldsymbol{\theta}_{-i,t}\right] d\theta_i, \end{aligned}$$

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

$$-\frac{\partial p\left(\mathbf{x}^i\right)}{\partial x_i} = \sum_{j \setminus i} -\max\left\{0, \frac{\partial U_j\left(\mathbf{x}^i\right)/\partial x_i}{\rho_j \Delta U_j(\mathbf{x}^*)}\right\} G_j'\left(\widetilde{\theta}_1\left(\mathbf{x}^i\right), ..., \widetilde{\theta}_{i-1}\left(\mathbf{x}^i\right), \widetilde{\theta}_{i+1}\left(\mathbf{x}^i\right), ...\widetilde{\theta}_n\left(\mathbf{x}^i\right)\right), \tag{12}$$

and, at the equilibrium, $\mathbf{x}^i = \mathbf{x}^*$,

$$\frac{\partial p(\mathbf{x}^*)}{\partial x_i} = \sum_{j \setminus i} \max\left\{0, \frac{\partial U_j(\mathbf{x}^*)/\partial x_i}{\rho_j \Delta U_j(\mathbf{x}^*)}\right\} G_j'(\mathbf{0}) = -\sum_{j \setminus i} \max\left\{0, \frac{\partial U_j(\mathbf{x}^*)/\partial x_i}{\rho_j \Delta U_j(\mathbf{x}^*)}\right\} f_j(0), \tag{13}$$

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

$$\left(1-p\left(\mathbf{x}^i\right)\right)U_i\left(\mathbf{x}^i\right)+p\left(\mathbf{x}^i\right)\left(1-\mathrm{E}\theta_{i,t}^R\rho_i\Delta\right)U_i\left(\mathbf{x}^*\right), \tag{14}$$

where $\mathrm{E}\theta_{i,t}^R$ 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^i, suppose i considers a small (marginal) reduction in x_i relative to x_i^*, given by $dx_i = x_i^i - x_i^* < 0$. If accepted, this gives i utility $U_i\left(\mathbf{x}^i\right) \approx U_i\left(\mathbf{x}^*\right) + dx_i \partial U_i\left(\mathbf{x}^*\right)/\partial x_i > U_i\left(\mathbf{x}^*\right)$, but it is rejected with probability

$$\frac{\partial p\left(\mathbf{x}^*\right)}{\partial x_i}dx_i = -\sum_{j\setminus i}\max\left\{0,\frac{\partial U_j\left(\mathbf{x}^*\right)/\partial x_i}{\rho_j\Delta U_j\left(\mathbf{x}^*\right)}\right\}dx_i f_j\left(0\right),$$

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., $\Pr\left(\theta_{j,t}\leq\widehat{\theta}_j\right)$ for $\widehat{\theta}_j \equiv \frac{\partial U_j\left(\mathbf{x}^i\right)/\partial x_i}{\rho_j\Delta U_j\left(\mathbf{x}^*\right)}\left|dx_i\right|$. Naturally, the probability that more than one party has such a small shock vanishes when $|dx_i|\to 0$ since f is assumed to have no mass point.

In combination, the reduction in x_i is not beneficial to i iff:

$$\begin{aligned}&\left(1-\frac{\partial p\left(x^*\right)}{\partial x_i}dx_i\right)\left(U_i\left(\mathbf{x}^*\right)+dx_i\frac{\partial U_i\left(\mathbf{x}^*\right)}{\partial x_i}\right)\\ -\sum_{j\setminus i}\max\left\{0,\frac{\partial U_j\left(\mathbf{x}^*\right)/\partial x_i}{\rho_j\Delta U_j\left(\mathbf{x}^*\right)}\right\}&dx_i f_j\left(0\right)U_i\left(\mathbf{x}^*\right)\left(1-E\left(\theta_{it}\mid\theta_{jt}\leq\widehat{\theta}_{j,t}\right)\rho_i\Delta\right)\leq \qquad U_i\left(\mathbf{x}^*\right),\end{aligned} \tag{15}$$

where (13) has been substituted into the second line of (15), and where $\mathrm{E}\left(\theta_{i,t}\mid\theta_{j,t}\leq\widehat{\theta}_j\right)$ follows from Bayes' rule:

$$\mathrm{E}\left(\theta_{i,t}\mid\theta_{j,t}\leq\widehat{\theta}_j\right)\equiv\frac{\int_0^{\widehat{\theta}_j}\int_{\Theta_{-j}}\theta_{i,t}f\left(\boldsymbol{\theta}_t\right)d\theta_j}{\int_0^{\widehat{\theta}_j}\int_{\Theta_{-j}}f\left(\boldsymbol{\theta}_t\right)d\theta_j},\ \text{and}\ \mathrm{E}\left(\theta_{i,t}\mid\theta_{j,t}=0\right)\equiv\lim_{dx_i\uparrow 0}\frac{\int_0^{\widehat{\theta}_j}\int_{\Theta_{-j}}\theta_{i,t}f\left(\boldsymbol{\theta}_t\right)d\theta_j}{\int_0^{\widehat{\theta}_j}\int_{\Theta_{-j}}f\left(\boldsymbol{\theta}_t\right)d\theta_j},$$

and, as already defined, $\Theta_{-j}\equiv\prod_{k\neq j}\left[0,\overline{\theta}_k\right]$ and $\widehat{\theta}_j\equiv\frac{\partial U_j(\mathbf{x}^*)/\partial x_i}{\rho_j\Delta U_j(\mathbf{x}^*)}\left|dx_i\right|$.

When both sides of (15) are divided by $|dx_i|$ and $dx_i\uparrow 0$, (15) can be rewritten as the first-order condition (11).

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

Proof of Theorem 1(ii).

A continuum of $\mathbf{x}^*$'s can satisfy the equilibrium condition in Theorem 1(i), since it is not necessary for i to improve an offer relative to $\mathbf{x}^*$ when $p\left(\mathbf{x}^*\right)=0$. The idea of the Small Trembles is to introduce trembles such that $p\left(\mathbf{x}^*\right)>0$ and thus we must check that i cannot benefit from marginally increasing *or* decreasing x_i^i from x_i^* to reduce $p\left(\mathbf{x}^i\right)$. With the Small Trembles, i will strictly benefit from $dx_i>0$ when (3) is strict, and thus it must hold with equality at $\mathbf{x}^*$.

The vector $\boldsymbol{\epsilon}_t$ is i.i.d. over time according to some cdf, $H\left(\cdot\right)$. (For simplicity, I omit the superscript k.) When j considers whether to accept $U_j\left(\mathbf{x}^i+\boldsymbol{\epsilon}_t\right)$, j faces the continuation value $V_j\left(\mathbf{x}^*\right)$ by rejecting, where $V_j\left(\mathbf{x}^*\right)$ takes into account that $\mathbf{x}^*$ may be rejected in the future (if the future $\epsilon_{i,t'}$'s are sufficiently small).[15] The shocks, combined with the possibility to reject, imply that $V_j\left(\mathbf{x}^*\right)>0$ even if $U_j\left(\mathbf{x}^*\right)=0$, so there is no need to assume $U_j\left(\mathbf{x}^*\right)>0\forall j$.

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

$$V_j\left(\mathbf{x}^*\right)=\mathrm{E}_{\boldsymbol{\epsilon}_{j,t}:(\boldsymbol{\epsilon},\theta)\in\Phi_A(\mathbf{x}^*)}\left(1-p\left(\mathbf{x}^*\right)\right)U_j\left(\mathbf{x}^*+\boldsymbol{\epsilon}_t\right)+p\left(\mathbf{x}^*\right)V_j\left(\mathbf{x}^*\right)\mathrm{E}_{\theta_{j,t}:(\boldsymbol{\epsilon},\theta)\in\Phi_R(\mathbf{x}^*)}\left(1-\theta_{j,t}\rho_j\Delta\right),$$

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

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

$$\left(1-p_{-j}\left(\mathbf{x}^i\right)\right)U_j\left(\mathbf{x}^i+\boldsymbol{\epsilon}_t\right)+p_{-j}\left(\mathbf{x}^i\right)\left(1-\rho_{j,t}\Delta\right)V_j\left(\mathbf{x}^*\right)<\left(1-\rho_{j,t}\Delta\right)V_j\left(\mathbf{x}^*\right)\Rightarrow$$

$$1-\theta_{j,t}\rho_j\Delta>\frac{U_j\left(\mathbf{x}^i+\boldsymbol{\epsilon}_t\right)}{V_j\left(\mathbf{x}^*\right)}\Rightarrow\theta_{j,t}<\widetilde{\theta}_j\left(\mathbf{x}^i\right)\equiv\frac{V_j\left(\mathbf{x}^*\right)-U_j\left(\mathbf{x}^i+\boldsymbol{\epsilon}_t\right)}{\rho_j\Delta V_j\left(\mathbf{x}^*\right)}.$$

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

$$\begin{aligned}1-p\left(\mathbf{x}^i\right) &= \int_{\boldsymbol{\epsilon}}G\left(\widetilde{\theta}_1\left(\mathbf{x}^i\right),...,\widetilde{\theta}_{i-1}\left(\mathbf{x}^i\right),\widetilde{\theta}_{i+1}\left(\mathbf{x}^i\right),...\widetilde{\theta}_n\left(\mathbf{x}^i\right)\right)dH\left(\boldsymbol{\epsilon}\right)\\ &\equiv \int_{\boldsymbol{\epsilon}}\int_0^{\overline{\theta}_i}\left[\int_{\widetilde{\theta}_1(\mathbf{x}^i)}^{\overline{\theta}_1}...\int_{\widetilde{\theta}_{i-1}(\mathbf{x}^i)}^{\overline{\theta}_{i-1}}\int_{\widetilde{\theta}_{i+1}(\mathbf{x}^i)}^{\overline{\theta}_{i+1}}...\int_{\widetilde{\theta}_n(\mathbf{x}^i)}^{\overline{\theta}_n}f\left(\boldsymbol{\theta}_t\right)d\boldsymbol{\theta}_{-i,t}\right]d\theta_i dH\left(\boldsymbol{\epsilon}\right)\Rightarrow\end{aligned}$$

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

The condition under which i does not benefit from a marginal change dx_i is given by the analogously modified (15),[16] but now, since this inequality is continuous at $dx_i = 0$, it must hold whether dx_i is positive or negative; it thus has to hold with equality; and it thus holds with equality regardless of whether i could benefit from $dx_i > 0$ or $dx_i < 0$ (so, we do not need the assumptions $\partial U_i(\cdot)/\partial x_j > 0$ for $j \neq i$ and < 0 for $j = i$).

When we let $\boldsymbol{\epsilon}$ vanish ($\mathrm{E}\left(\boldsymbol{\epsilon}_t^k\right)^2 \to \mathbf{0}$ when $k \to 0$), we get $p\left(\mathbf{x}^*\right) \to 0$ and $V_j\left(\mathbf{x}^*\right) \to U_j\left(\mathbf{x}^*\right)$, and then the condition simplifies to

$$-\frac{\partial U_i\left(\mathbf{x}^*\right)}{\partial x_i}=\sum_{j\setminus i}\frac{\partial U_j\left(\mathbf{x}^*\right)/\partial x_i}{\rho_j\Delta U_j\left(\mathbf{x}^*\right)}f_j\left(0\right)\mathrm{E}\left(\theta_{i,t}\mid\theta_{j,t}=0\right)\rho_i\Delta U_i\left(\mathbf{x}^*\right),$$

which is the first-order condition of

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

when $\frac{w_j^i}{w_i^i}=\frac{\rho_i}{\rho_j}f_j\left(0\right)\mathrm{E}(\theta_{i,t}\mid\theta_{j,t}=0)$, $\forall j\neq i$. *QED*

The other proofs are rather standard and will be added to the paper soon.

[16] The modified version of (15) can be written as:

$$\begin{aligned}&\left(1-p\left(\mathbf{x}^*\right)-\frac{\partial p\left(\mathbf{x}^*\right)}{\partial x_i}dx_i\right)\mathrm{E}_{\boldsymbol{\epsilon}_{i,t}:(\boldsymbol{\epsilon},\theta)\in\Phi_A(\mathbf{x}^*)}\left(U_i\left(\mathbf{x}^*+\boldsymbol{\epsilon}\right)+\frac{\partial U_i\left(\mathbf{x}^*+\boldsymbol{\epsilon}\right)}{\partial x_i}dx_i\right)\\ &+\int_{\epsilon}\sum_{j\setminus i}\left[\frac{\partial U_j\left(\mathbf{x}^*+\boldsymbol{\epsilon}\right)/\partial x_i}{\rho_j\Delta V_j\left(\mathbf{x}^*\right)}dx_i\int_0^{\overline{\theta}_i}G_j'\left(\frac{V_1\left(\mathbf{x}^*\right)-U_1\left(\mathbf{x}^*+\boldsymbol{\epsilon}\right)}{\rho_1\Delta V_1\left(\mathbf{x}^*\right)},\frac{V_2\left(\mathbf{x}^*\right)-U_2\left(\mathbf{x}^*+\boldsymbol{\epsilon}\right)}{\rho_2\Delta V_2\left(\mathbf{x}^*\right)},...;\theta_i\right)d\theta_i\right].\\ &U_i\left(\mathbf{x}^*\right)\mathrm{E}_{\theta_{i,t}:(\boldsymbol{\epsilon},\theta)\in\Phi_R(\mathbf{x}^*)}\left(1-\theta_{i,t}\rho_i\Delta\right)\leq U_i\left(\mathbf{x}^*\right).\end{aligned}$$

References [Preliminary]

Aldy, Joseph E., Barrett, Scott and Stavins, Robert N. (2003): "Thirteen plus one: a comparison of global climate policy architectures," *Climate Policy* 3(4): 373-397.

Andersson, Ola; Argenton, Cédric and Weibull, Jörgen W. (2017): "Robustness to Strategic Uncertainty in the Nash Demand Game," forthcoming, *Mathematical Social Sciences.*

Abreu, Dilip and Gul, Faruk (2000): "Bargaining and Reputation," *Econometrica* 68(1): 85-117.

Abreu, Dilip and Pearce, David (2007): "Bargaining, Reputation, and Equilibrium Selection in Repeated Games with Contracts," *Econometrica* 75(3): 653-710.

Abreu, Dilip and Pearce, David (2015): "A Dynamic Reinterpretation of Nash Bargaining With Endogenous Threats," *Econometrica* 83(4): 1641-1655.

Barrett, Scott (1994): "Self-enforcing international environmental agreements," *Oxford Economic Papers*, 46, p. 878-94.

Barrett, Scott and Dannenberg, Astrid (2016): "An experimental investigation into 'pledge and review' in climate negotiations," *Climatic Change* 138(1): 339-351.

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

Binmore, Ken; Rubinstein, Ariel and Wolinsky, Asher (1986): "The Nash Bargaining Solutin in Economic Modelling," *The RAND Journal of Economics* 17(2): 176-188.

Binmore, Ken (1987): "Nash Bargaining Theory (II)," in *The Economics of Bargaining*, ed. by K. Binmore and P. Dasgupta. Cambridge: Basil Blackwell.

Binmore, Ken; Osborne, Martin J. and Rubinstein, Ariel (1992): "Non-Cooperative Models of Bargaining," *Handbook of Game Theory*, Volume 1, Aumann, R. J.; Hart, S (ed.), Elsevier Science Publishers.

Britz, Volker; Herings, P. Jean-Jacques and Predtetchinski, Arkadi (2010): "Non-cooperative Support for the Asymmetric Nash Bargaining Solution," *Journal of Economic Theory* 145: 1951-1967.

Carlsson, Hans (1991): "A Bargaining Model Where Parties Make Errors," *Econometrica* 59(5): 1487-1496.

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

Cho, In-Koo and Matsui, Akihiko (2013): "Search Theory, Competitive Equilibrium and the Nash Bargaining Solution," *Journal of Economic Theory* 148(4): 1659-1688.

d'Aspremont, Claude., Jacquemin, Alexis, Gabszewicz, Jean jaskold and Weymark, John A. (1983): "On the stability of collusive price leadership," *The Canadian Journal of Economics* 16(1): 17–25.

Dutta, Prajit K. and Radner, Roy (2004): "Self-enforcing climate-change treaties," *Proceedings of the National Academy of Science*, 101: 4746-51.

Finus, Michael and Maus, Stefan (2008): "Modesty may pay!" *Journal of Public Economic Theory* 10: 801–26.

Harsanyi, John and Selten, Reinhard (1972): "A Generalized Nash Solution for Two-Person Bargaining Games with Incomplete Information," *Management Science* 18(5) part 2: 80-106.

Harstad, Bård (2016): "The Dynamics of Climate Agreements," *Journal of the European Economic Association.*

Harstad, Bård, Lancia, Francesco and Russo, Alessia (2018): "Compliance Technology and Self-Enforcing Agreements," mimeo, University of Oslo.

Hoel, Michael (1992): "International environmental conventions: the case of uniform reductions of emissions," *Environmental and Resource Economics* 2(2): 141-159.

Howard, J. V. (1992): "A social choice rule and its implementation in perfect equilibrium," *Journal of Economic Theory* 56(1): 142-159.

Kalai, Ehud (1977): "Non-symmetric Nash Solutions and Replication of 2-Person Bargaining," *International Journal of Game Theory* 6(3): 129-133.

Kambe, Shinsuke (2000): "Bargaining with Imperfect Commitment," *Games and Economic Behavior* 28: 217-237.

Kawamori, Tomohiko (2014): "A Non-Cooperative Foundation of the Asymmetric Nash Bargaining Solution," *Journal of Mathematical Economics* 52: 12-15.

Kolstad, Charles D. and Toman, Michael (2005): "The Economics of Climate Policy," *Handbook of Environmental Economics* 3: 1562-93.

Golosov, M., Hassler, J., Krusell P. and A. Tsyvinski (2014): "Optimal Taxes on Fossil Fuel in General Equilibrium," *Econometrica* 82(1): 41-88.

Laurelle, Annick and Valenciano, Federico (2008): "Non-Cooperative Foundations of Bargaining Power in Committees and the Shapley-Shubik Index," *Games and Economic Behavior* 63: 341-353

Miyakawa, Toshiji (2008): "Note on the Equal Split Solution in an n-Person Non-Cooperative Bargaining Game," *Mathematical Social Sciences* 55(3): 281-291.

Nash, John (1950): "The Bargaining Problem," *Econometrica* 18: 155-162.

Nash, John (1953): "Two-Person Cooperative Games," *Econometrica* 21(1): 128-140.

Nordhaus, William D. (2006): "After Kyoto: Alternative Mechanisms to Control Global Warming," *American Economic Review* 96(2): 31–4.

Osborne, Martin J. and Rubinstein, Ariel (1990): *Bargaining and Markets*, Academic Press.

Rey, Debraj and Vohra, Rajiv (2001): "Coalitional Power and Public Goods," *Journal of Political Economy* 109 (6): 1355-84.

Roth, Alvin E. (1979): "Proportional Solutions to the Bargaining Problem," *Econometrica* 47(3): 775-778.

Rubinstein, Ariel (1982): "Perfect Equilibrium in a Bargaining Model," *Econometrica* 50(1): 97-109.

Sutton, John (1986): "Non-Cooperative Bargaining Theory: An Introduction," *The Review of Economic Studies* 53(5): 709-724.