# International Climate Change Mitigation, Information Dissemination, and Domestic Politics*

Justin Melnick†

February 13, 2024

**Abstract**

The 2016 Paris Agreement sought to overcome previous international climate cooperation failures by allowing countries to propose their own emissions reductions targets; international cooperation served to elicit and disseminate information about these reports. Previous work suggests that these institutional features, voluntary commitments and publicity of reports, should deepen cooperation. I develop a formal model that characterizes optimal cooperation within an agreement like Paris that allows leaders to voluntarily reveal their nation's willingness to invest in mitigation in the shadow of domestic electoral politics. I document that, when leaders' private information about their willingness to mitigate is electorally relevant information, the presence of an agreement like Paris leads to less ambitious climate agendas. The result arises because the transparency of reports facilitated by the international organization provides voters with information needed to resolve a domestic electoral selection problem, which disincentivizes leaders to invest in mitigation efforts.

Word Count: 9493 (text) + 1093 (references) + 2135 (appendix) = 12721 (total)

*I am grateful to Pat Egan, Sanford Gordon, Layna Mosley, Peter Rosendorff, Tara Slough, Alastair Smith, Aleksandra Conevska, Kun Heo, the NYU IR Workshop, NYU Dissertation Seminar, the Harvard Political Economy Workshop, the SWIPER reading group, and EITM 2023 for insightful comments and feedback.

†Ph.D. Candidate, Department of Politics, New York University. Contact: melnickj@nyu.edu and justinmelnick.github.io.

Prior to the adoption of the Paris Agreement in 2016, scholars and policymakers lamented that the international community had reached the "climate dead end," failing to develop enduring international institutions to address the threat of climate change (Stern 2007; Keohane and Oppenheimer 2016). Paris's predecessor, the 1997 Kyoto Protocol, collapsed because its institutional structure was not conducive to deep international cooperation (Barrett 2003; Depledge 2006; Hoffmann 2008). Notably, Kyoto imposed uniform, "top-down," legally binding emissions reductions targets onto industrialized nations and levied fines for noncompliance, ultimately leading several countries to delay ratification and others to withdraw from the agreement altogether (Victor 2011; Keohane and Oppenheimer 2016); Kyoto ultimately fell victim to common political instabilities that perpetuate the underprovision of global public goods (Keohane and Victor 2016). By contrast, Paris seeks to move beyond the climate dead end by reshaping leaders' incentives to invest in costly mitigation efforts. The new agreement opts for a "bottom-up" approach, allowing countries to propose their own emissions reductions targets, implemented via national investments into mitigation policy. Additionally, the agreement contains no formal enforcement mechanism at the international level: the institution's role is simply to collect and disseminate information about nations' willingness to mitigate.

How does the revelation of information by an international institution affect the prospects for cooperation? In this paper, I develop a formal model of international cooperation with domestic politics within an agreement like Paris in which leaders voluntarily reveal their willingness to invest in mitigation to an international organization that makes such reports public. I consider how these two novel institutional features of the Paris Agreement – voluntary commitments and publicity of reports – shape leaders' incentives to report and subsequently invest in costly mitigation efforts. Conventional wisdom on the structure of international agreements would contend that these features should increase cooperation due to increased flexibility (von Stein 2008; Keohane and Victor 2011) and transparency (Bodansky 2016; Jacquet and Jamieson 2016; Dannenberg et al. 2023). Contrarily, I demonstrate that these features have deleterious effects on leaders' investments into mitigation: Paris's institutional design inspires less ambitious mitigation efforts through its effects on domestic politics.

To study how Paris's informational effects alter the prospects for climate action, I compare equilibria of two models. In the first, I study the interaction between a leader and a voter without any overarching international institution. Mitigation efforts induce political costs for leaders because of their distributive implications (Aklin and Mildenberger 2020), and leaders vary in their willingness to shoulder these costs. These marginal costs of investing in mitigation, a leader's "type," determine effort exerted into climate action. The voter observes an imperfect signal of effort, the implementation of mitigation policy, which is informative of leader type. Policy implementation forms the basis of an accountability relationship between the leader and the voter (cf. Ashworth, Bueno de Mesquita and Friedenberg 2018; Gailmard and Patty 2019; Balcazar and Kennard 2023); the leader's reelection prospects are thus conditioned on observable policy outcomes. Given evidence that the general public supports diffuse redistribution to combat climate change (Gaikwad, Genovese and Tingley 2022), higher realizations of the voter's signal represent more "successful" mitigation policy implementation, which on average reflects a leader who invested greater efforts into mitigation. Leaders therefore find it politically advantageous to invest in mitigation efforts as a "good" signal to voters.

The second model introduces an international organization (IO) seeking to maximize global mitigation efforts. To capture the bottom-up flavor of Paris, I consider a setting in which leaders report their type to the IO, representing their "nationally determined contribution." The IO makes these reports public and subsequently recommends a level of effort for each leader to pursue. I follow extant approaches (Harrison and Lagunoff 2017; Slechten 2020; McAllister and Schnakenberg 2022) and analyze the optimal effort recommendations that the IO could make given leaders' political constraints. The core insight is that, because information about leader type is also domestically relevant for voters, Paris's design of aggregating and disseminating information detracts from leaders' incentives to exert effort into mitigation rather than enhance them. Under Paris, leaders' efforts into mitigation are lower in a world without international cooperation.

Paris detracts from mitigation efforts because of its informational effects on domestic politics. The direct provision of information about leader type breaks the accountability relationship between leaders and voters because the IO provides information to voters that resolves uncertainty about leader type. Consequently, voter assessments of a leader's willingness to mitigate are no longer conditioned on leader effort, hence leaders' electoral incentives to exert effort dissipate. Thus, the transparency of the Paris Agreement, often heralded as one of the central features of the institution to foster more ambitious climate action (Falkner 2016), unravels leaders' incentives to invest in costly mitigation efforts.

The theory in advanced in this paper relies on an informational mechanism to establish the domestic microfoundations of leaders' efforts to address the effects of climate change. This approach is most closely related to the theoretical argument made by Balcazar and Kennard (2023), who demonstrate how climate shocks can provide information about one's political environment. This model abstracts away from the direct environmental impacts of climate change, but documents how the implementation of climate *policy* can be informative of electorally relevant leader traits through a similar accountability mechanism.

This paper makes substantive and theoretical contributions, speaking to several literatures in international politics and political economy. Principally, I contribute to work elucidating international organizations' role of disseminating information to enhance cooperation (Keohane 1984; Dai 2005); in particular, I focus on the international provision of information to domestic audiences. Scholars have demonstrated that international institutions may motivate good governance by disseminating information about government performance (Kelley and Simmons 2015; 2019) with the hopes of disciplining government behavior (Besley and Burgess 2002; Hollyer, Rosendorff and Vreeland 2015). These studies contend that greater information provision will lead to better governance because information should strengthen the accountability channel between leaders and their publics; however, this paper joins other work that demonstrates how information from international institutions can worsen governance outcomes by distorting leaders' incentives within these domestic relationships (e.g., Hollyer and Rosendorff 2011; Bisbee et al. 2019).

Moreover, this paper has implications for the design of climate governance institutions. By recasting the problem of international climate cooperation as a bottom-up endeavor, Paris has redefined the incentives for leaders to interact with international institutions. Compared to extant agreements like the Kyoto Protocol, Paris places the onus on leaders and their domestic publics to determine the political feasibility of mitigation efforts rather than striving to comply with legally binding international reduction targets (Falkner 2016; Harstad 2023). This model identifies how two critical features of contemporary international climate governance – voluntary contributions and the publicity of those contributions – can generate deleterious international outcomes when situated within leaders' domestic accountability contexts.

This paper also contributes to the political economy literature on transparency in agency relationships. Formal models of transparency show that transparency can worsen governance outcomes within accountability channels (Prat 2005; Fox and Van Weelden 2012). Besley (2006) shows that the direct provision of information decreases accountability. This model reaffirms this finding. In the domestic context, the desire to remain in office motivates leaders to exert greater effort into mitigation policy, as doing so is more likely to realize more successful policy outcomes, which reflect desirably on type and thus increase reelection prospects. However, Paris's transparency at the international level unravels this accountability relationship. By publicizing reports about leader type, the agreement provides voters with exactly what they need to assess their leaders, who in turn no longer have electoral incentives to deepen cooperation.

Finally, this model connects to extant work in formal theory that examines climate change cooperation as a mechanism design problem (Harrison and Lagunoff 2017; Slechten 2020; McAllister and Schnakenberg 2022). Previous work assumes leaders face exogenous costs for noncompliance and that an international organization with sufficient enforcement power exists to impose these costs. Such explanations both lack theoretical microfoundations for how or why leaders might face costs of noncompliance, and fail to recognize enforcement capabilities in the international climate governance realm are minimal. Indeed, the Paris Agreement's only enforcement mechanism is "naming and shaming" through the dissemination of information. I innovate on this framework by microfounding leaders' incentives to cooperate within a model of domestic politics, and in so doing demonstrate new obstacles to cooperation.

The paper's analysis proceeds by comparing two models. In the first, leaders invest in mitigation policies in the shadow of domestic politics without any overarching international institution (the domestic politics game). This microfounds the insight that leaders may be incentivized to invest effort into mitigation for domestic electoral considerations. In the second, I introduce cooperation whereby an international organization makes effort recommendations for leaders to carry out and characterize the optimal recommendations given leaders' political constraints (the international cooperation game). I demonstrate how cooperation unravels through the intersection of public reporting and domestic electoral selection.

## Domestic Politics Game

I begin by considering a model in which leaders exert effort into reducing emissions without any structured international cooperation.

### Setup

Consider a strategic interaction between $n > 2$ countries that engage in policymaking on the mitigation of carbon emissions. In each country (indexed by $i$) there is a leader and a representative voter. With slight abuse of notation, I will often refer to the leader of country $i$ as leader $i$. Leader $i$ has a private type, $\theta_i \in \{\underline{\theta}\ \overline{\theta}\}$ with $0 < \underline{\theta} < \overline{\theta}$. Let the common prior be $P(\theta_i = \underline{\theta}) = q$. Only leader $i$ knows her type; all voters and all other leaders only know the prior. Leaders' types represent their "willingness to mitigate," or the extent to which they find it costly to invest in mitigation efforts. Leaders of type $\underline{\theta}$ have lower costs and are therefore more willing to invest in mitigation, while leaders of type $\overline{\theta}$ face greater costs and are subsequently less willing to mitigate. I will also refer to leaders of type $\underline{\theta}$ as "congruent," and leaders of type $\overline{\theta}$ as "incongruent." More will be explained below on the interpretation of leader type.

Given their types, leaders choose a policy $a_i \in [0, \omega]$. These policies represent effort allocated toward mitigation investments, where $\omega > 0$ is some maximum feasible effort level. A higher value of $a_i$ is consistent with more ambitious effort and thus a greater commitment to mitigation on behalf of country $i$. When investing effort, leaders care about the tradeoff between decreasing domestic emissions, thus incurring the costs of mitigation, and free riding off of the efforts of other nations. By taking fewer steps toward abatement, or investing in a lower $a_i$, leader $i$ onshores less of the costs required to engage in global mitigation efforts. To capture these considerations, I endow leader $i$ with a utility function over mitigation efforts with the following form:[1]

$$u(a_i, A; \theta_i) = A - \frac{\theta_i}{2} a_i^2,$$

where $A = \sum a_i$ is global effort to reduce emissions. To avoid corner solutions, I make the following restrictions on the parameter space: $0 < \frac{n}{\omega} \leq \underline{\theta} < \overline{\theta} \leq n < \omega$.

The utility function captures several key considerations of the climate mitigation problem. First, leaders benefit from aggregate reductions. Allocating costly effort toward reducing emissions is akin to providing a public good (Kennard and Schnakenberg 2023); each leader is better off when nations work harder to increase their mitigation programs. However, mitigation efforts are costly. This establishes the temptation for leaders to free ride off of the contributions of others. Third, as mentioned, leaders vary in their willingness to implement mitigation measures, parameterized as the marginal costs of exerting effort, their type $\theta_i$.

In addition to engaging with other nations, leaders are accountable to domestic publics when developing mitigation strategies. The voter sees a noisy signal of effort,

$$K_i = a_i + \varepsilon_i,$$

where $\varepsilon_i \sim N(0, \frac{1}{\beta})$. I assume that $\beta \leq \frac{2\gamma\overline{\theta}\sqrt{2\pi e}}{\Psi}$.[2]

The signal $K_i$ literally implies that voters have imperfect information about leader effort, but could also represent the net value of climate policy, or the "success" in implementing mitigation policy, at the time of the election. Uncertainty in the mapping from effort to policy outcomes may also represent leaders' inability to forecast how policy may affect a pivotal voter (Gazmararian and Tingley 2023). Higher values of the signal are more likely to reflect more ambitious effort levels. Based on the realization of the signal, voter $i$ determines whether to retain leader $i$ or replace her, choosing $\rho_i \in \{0, 1\}$ where $\rho_i = 1$ denotes that the voter retains the leader and $\rho_i = 0$ denotes that he replaces her.

Two elements comprise the voter's payoff. A voter receives benefits from having a congruent leader in office, but, in addition to the selection problem, he also has a predisposed bias toward the incumbent leader, which represents the value of the incumbent on all other dimensions besides the implementation of mitigation

[1] This specific functional form follows Harstad (2023) but is also consistent with the assumptions placed on utility in McAllister and Schnakenberg (2022).

[2] This is a sufficient condition to ensure that the equilibrium effort for the incongruent leader is a maximum. See proof of Proposition 1 in the appendix.

strategies. Denote bias as $y_i \sim G(\cdot)$, where $G(\cdot)$ is a cumulative distribution function satisfying the monotone likelihood ratio property. For analytical simplicity, I will work with $y_i \sim U[-\gamma,\ \gamma]$. To easily parameterize the voter's preferences for congruence, suppose that the voter gets utility 1 from having a congruent leader in power and 0 otherwise. If the voter replaces the incumbent leader, her replacement is of type $\theta_C$ such that $P(\theta_C = \underline{\theta}) = q$. The voter's utility function is thus

$$u(\rho_i; y_i) = \rho_i \Big( \mathbb{1}_{\theta_i = \underline{\theta}} + y_i \Big) + (1 - \rho_i) \mathbb{1}_{\theta_C = \underline{\theta}}.$$

Note that the voter does not have have explicit, material preferences over the implementation of climate policy, nor does he bear any costs of complying with mitigation policies. I simply require that the voter gleans some electorally relevant information through the implementation of climate policy.

The leader's payoff comprises of both utility over mitigation policy $u(a_i, A; \theta_i)$ and a benefit from remaining in power. If the voter retains the leader, the leader enjoys the value of holding office, $\Psi > 0$. The leader's payoff is thus

$$U(a_i, \rho_i; \theta_i) = u(a_i, A; \theta_i) + \rho_i \Psi.$$

The timing of the game is as follows:

1. Types $\theta_i$ are revealed to incumbent leaders.
2. Leaders simultaneously choose effort $a_i$.
3. Voters observe the signal $K_i$ and bias $y_i$ and retain or replace their leaders.
4. Payoffs realized. Game ends.

I examine Perfect Bayesian Equilibria. A leader's strategy maps type into an effort level, $\sigma_i : \{\underline{\theta}, \overline{\theta}\} \to [0, \omega]$. The voter's strategy maps the realization of the signal and the valence shock into a retention rule, $\rho_i : \mathbb{R} \times [-\gamma,\ \gamma] \to \{0, 1\}$. The voter's retention rule is sequentially rational given conjectures about the leader's policy choice (since it is imperfectly observed) and beliefs about the leader's type. The leader's effort choices are sequentially rational with the voter's retention rule. Beliefs about leader type are determined by Bayes's rule.

## Comments on the Model

The domestic politics game makes several simplifying assumptions that warrant further explanation.

**The leader and her incentives.** Leaders vary in their marginal costs of exerting effort into mitigation, which is their private type. This cost captures three considerations. First, formulating climate change policies may have political consequences for leaders because these policies have winners and losers (Aklin and Mildenberger 2020). When investing in mitigation strategies, leaders anticipate potential opposition, which may be politically costly (Stokes 2016; Breetz, Mildenberger and Stokes 2018); some leaders are more willing than others to incur these costs. Second, leaders vary in their *ex ante* ideological proclivity toward implementing mitigation policy, which may stem from engagements with pro- or anti-climate interests (Mildenberger 2020; Stokes 2020). Third, mitigation policy may carry opportunity costs in the form of forsaking other policy agendas.

Previous studies have operationalized leaders' private information as "national capabilities" to mitigate or a country's marginal cost of abatement (McAllister and Schnakenberg 2022). National capabilities to mitigate climate policy are a function of the domestic political environment. Thus, a nation's marginal cost of abatement should be dependent on a political leader's willingness to pursue policies that would advance mitigation efforts. For example, the United States' ability to mitigate has changed over time with leader turnover, with variance in top political leaders driving the country's climate agenda. President Obama displayed high willingness to mitigate – endorsing the Waxman-Markey bill which ultimately could not withstand a Senate filibuster – while President Trump retrenched America's capabilities to mitigate by opposing any legislation that would address the effects of climate change. By contrast, President Biden's signing of the Inflation Reduction Act signaled a high willingness to rectify the Trump Administration's antipathy toward climate mitigation policy.

**The voter and his incentives.** As is common in political agency models, I consider a representative voter who faces a selection problem with respect to leader type. Voters prefer congruent leaders to incongruent leaders, and thus prefer "good" policy implementation over "bad" policy implementation, insofar as it aids in selecting leaders who are more likely to exert effort toward mitigation. While the voter's preference for congruence is assumed into the payoffs, this could be microfounded with a second period of post-election policymaking; the results of such an extended model are not qualitatively different. Additional support for this assumption comes from the idea that all else equal, voters would prefer the implementation of policies addressing climate change (Flynn et al. 2021) or that the general public would be supportive of broad-based compensatory mechanisms to pursue mitigation (Gaikwad, Genovese and Tingley 2022).

Recent syntheses of the effects of climate policy on public opinion identify three broad areas in which policymaking can affect political attitudes and behavior: the visibility of an issue, an individual's prioritization of an issue, and the public's understanding of policy benefits and costs (Gazmararian, Mildenberger and Tingley 2023). The model captures each of these concepts in a stylized manner. The visibility of the issue may also correspond with the policy's informativeness, or the extent to which voters can learn about electorally important information through the implementation of climate policy. This is parameterized through $\beta$, the precision of the voter's signal of the leader's effort investments into mitigation.[3] An individual's prioritization of the issue maps to $\gamma$, the support of the valence shock, which allows the voter to adjudicate between the value of selecting congruent leaders and the value of the incumbent on other electorally salient dimensions of policy.[4] Finally, the tradeoff between the costs and benefits of policy may be understood as the observed value of the voter's signal, $K_i$, which represents the net value of the policy to the voter at the time of the election.

It is a feature of the model, not a bug, that the voter's incentives are highly stylized. Introducing greater complexity like imposing costs on voters for mitigation policies (cf. Bechtel and Scheve 2013; Ansolabehere and Konisky 2014) might weaken the accountability relationship, but would not alter how information about leader type is transmitted to voters, thus leaving the flavor of the posited mechanism intact. Moreover, that voters do not pay for mitigation policy reflects the idea that the redistributive consequences generated by climate policy are borne out by *leaders*, who may face electoral or other political costs for their actions, not voters themselves.

## Analysis

I now build intuition for the domestic informational mechanism that undergirds leaders' choices of mitigation efforts. Because the voter is assumed to prefer congruent leaders, he adopts a straightforward retention rule in which voter $i$ retains leader $i$ if and only if his posterior belief about the leader's congruence, as well as his bias, is greater than the probability that her replacement is congruent,

$$\mu(K_i) + y_i \geq q,$$

[3]As $\beta \to \infty$, voters receive a perfect signal of effort (and subsequently of type). Conversely, as $\beta \to 0$, the signal is completely uninformative, making it difficult for voters to select congruent leaders from the realization of policy outcomes.

[4]A high value of $\gamma$ means that the voter's bias toward the incumbent may be very large (in either direction), implying that the selection of congruent leaders provides little value added to a voter's electoral choice. If $\gamma > q$ or $-\gamma < -1 + q$, then there are realizations of the valence shock such that the incumbent leader will win or lose the election regardless of the effort invested into climate policy, reflecting a case where mitigation implementation is not a salient issue to voters.

where $\mu(K_i) = P(\underline{\theta}|K_i)$ is the voter's posterior belief about the leader's congruence given the observed value of the signal.

Since efforts are not perfectly observed, the voter in each country needs to have conjectures about the abatement policies chosen by each leader-type. Denote these by $(\hat{a}(\underline{\theta}),\ \hat{a}(\overline{\theta}))$. The voter retains the incumbent whenever the realization of mitigation policy is greater than or equal to some threshold, $K_i \geq \hat{K}$. Conditional on some value of the valence shock $y_i$, the voter is thus exactly indifferent between retaining the incumbent leader and replacing her when

$$K(y_i) = \frac{\hat{a}(\underline{\theta}) + \hat{a}(\overline{\theta})}{2} + \frac{\log\left(\frac{(1-q)(q-y_i)}{q(1-q+y_i)}\right)}{\beta(\hat{a}(\underline{\theta}) - \hat{a}(\overline{\theta}))} \ \equiv\ \hat{K}(y_i).$$

This equality provides a relationship between the value of the signal that the voter needs to observe in order to retain the incumbent and a realized value of the voter's valence shock; the voter's bias $y_i$ affects the way he parses information about congruence. Increasing bias $y_i$ decreases the relative value of selecting congruent leaders and therefore creates a more permissive cutoff rule by which voters assess leader type. If $y_i > q$, then the voter's bias toward the incumbent dwarfs the value added of the signal so that $K(y_i) \to -\infty$, and the voter would retain the incumbent for any value of $K_i$. Leaders who are already ingratiated with the voter thus survive in office regardless of how their effort into mitigation translates into climate policy. Conversely, if $y_i < -1 + q$, then his bias against the incumbent leader is large and $K(y_i) \to \infty$, meaning the leader would never be retained regardless of the signal.

Given the voter's cutoff rule, the probability that the leader survives in office can be recovered. Integrating out the valence shock, the probability that a leader of type $\theta_i$ is reelected is therefore

$$\frac{1}{2\gamma}\int_{-1+q}^{q} \Phi(\sqrt{\beta}(a_i - \hat{K}(y_i)))\ dy + \frac{\gamma - q}{2\gamma}.$$

The first term represents the chance that the leader is reelected as a function of her investment into mitigation policy. The second term is not a function of this effort, because there is some probability that the leader is reelected because she is a particularly high-valence leader. A leader of type $\theta_i$ therefore has an expected utility of

$$EU_i(a_i, a_{-i}; \theta_i) = a_i + \sum_{\theta_{-i}} a_{-i}(\theta_{-i})P(\theta_{-i}) - \frac{\theta_i}{2}a_i^2 + \Big[\int_{-1+q}^{q} \Phi(\sqrt{\beta}(a_i - \hat{K}(y_i)))\ dy + \gamma - q\Big]\frac{\Psi}{2\gamma}.$$

Leaders care about their utility derived from pure policy benefits from mitigating but also about the value of remaining in office. Notice immediately that if electoral incentives are irrelevant ($\Psi = 0$), this maximization resembles a classic problem of public goods provision; without electoral incentives, leaders simply invest in mitigation policy according to what maximizes their tradeoff between the value of committing to reducing emissions and the personal costs of doing so. In such a case, leaders would choose their "ideal" efforts:

$$\tilde{a}(\theta_i) = \frac{1}{\theta_i}.$$

The ideal effort level serves as a convenient benchmark because it represents the level of effort that leaders would exert if the accountability mechanism were blackboxed but leaders factored in other possible costs associated with implementing mitigation strategies. This includes both the domestic costs like possible redistributive consequences from policy as well as the value of the global collective action problem of exerting effort to reduce emissions.

Intuitively, leaders with greater willingness to invest in climate mitigation will exert greater effort, $\tilde{a}(\underline{\theta}) > \tilde{a}(\overline{\theta})$. However, leaders also generally care about holding office. When choosing optimal mitigation efforts, leaders also consider how their policy choices might reveal information about their type, and subsequently their electoral prospects. Leader-type $\theta_i$'s effort choice satisfies the following first-order condition (applying the refinement that the voter's beliefs about effort choices are correct in equilibrium):

$$\underbrace{\theta_i a^*(\theta_i)}_{\substack{\text{marginal willingness to} \\ \text{exert effort}}} = \underbrace{1}_{\substack{\text{marginal global benefit of} \\ \text{exerting effort}}} + \underbrace{\frac{\sqrt{\beta}\Psi}{2\gamma} \int_{-1+q}^{q} \phi\Big(\sqrt{\beta}(a^*(\theta_i) - \hat{K}^*(y_i))\Big)\, dy}_{\substack{\text{marginal benefit of} \\ \text{office-holding}}}.$$

Leaders balance the marginal costs of exerting effort into mitigation policy with two forces. Leaders care about the benefits that mitigation efforts bring, thus contributing to the global public good, and leaders exert effort as a means of surviving in office by signaling congruence to the voter. Manipulating this first-order condition yields several insights. First, leaders always behave more ambitiously when abatement benefits them electorally. Compared to the ideal level of effort, leaders always increase their commitments to reducing emissions in the presence of electoral incentives. Formally, if $a_i = \frac{1}{\theta_i}$, the above first-order condition becomes $\frac{\sqrt{\beta}\Psi}{2\gamma} \int_{-1+q}^{q} \phi\Big(\sqrt{\beta}(a^*(\theta_i) - \hat{K}^*(y_i))\Big)\, dy > 0$, which means that the ideal investment is less than the utility-maximizing effort, $\tilde{a}(\theta_i) < a^*(\theta_i)$. Second, congruent leaders still abate more than incongruent leaders. But now, the voter acts as a counter on the leader's temptations to pursue their less ambitious, ideal commitments: exerting lower effort on average will signal to the voter that the leader is incongruent.

In equilibrium, the voter's signal follows the distribution $K_i \sim N(a^*(\theta_i), \frac{1}{\beta})$, so how leaders exert effort in equilibrium defines the *distribution* of possible policy outcomes that the voter can observe. This is illustrated in Figure 1. In the figure, the purple solid line is the distribution of $K_i$ for a congruent leader, and the grey dashed line is the distribution of $K_i$ for an incongruent leader. It is evident from the figure that the congruent leader invests more effort into mitigation strategies because the solid curve has greater density toward higher values of $K_i$. The black dotted line represents the voter's equilibrium cutoff $\hat{K}(a^*(\underline{\theta}),\ a^*(\overline{\theta}))$ (drawn here with $y_i = 0$). Any draw from these distributions to the right of this line represents a signal for which the voter would retain the leader, while any draw to the left means the voter replaces the leader.

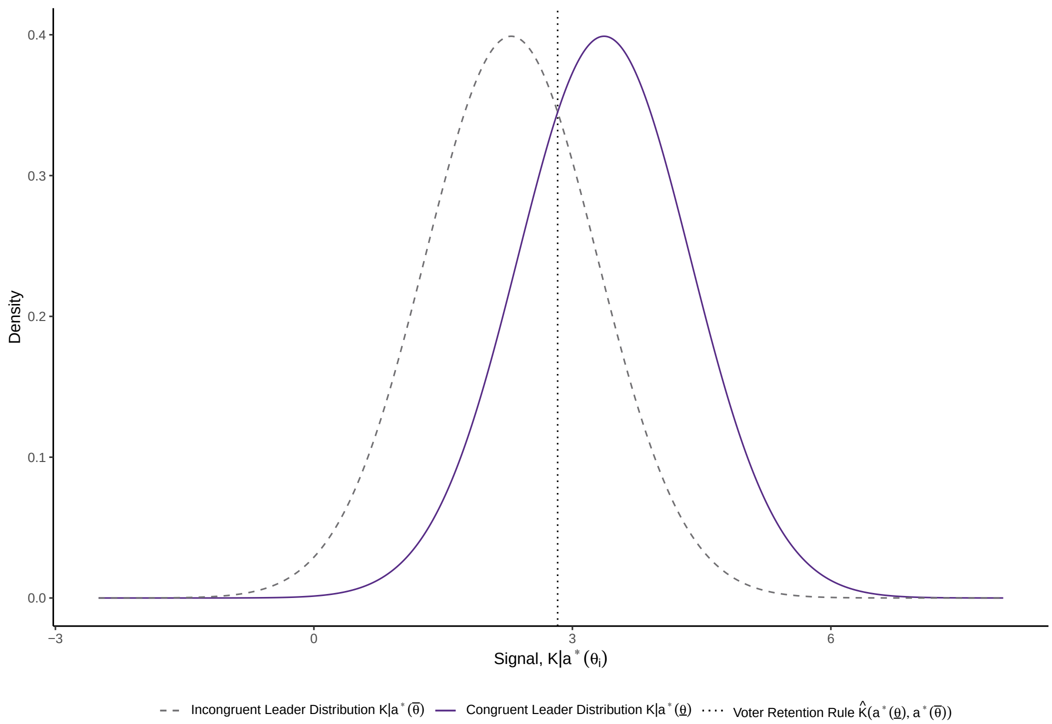


Figure 1: Equilibrium Signal Distributions in the Domestic Politics Game

In equilibrium, the voter's selection problem is resolved in expectation, but he does not perfectly screen the congruence of the leader. That is, the voter sometimes punishes congruent leaders or rewards incongruent leaders because of extreme values of the signal $K_i$, or because of the realization of the valence shock. However, the probability that a congruent leader survives is always greater than the probability that an incongruent leaders survives. Thus, on average, the voter makes the "right" decision. These probabilities can be seen graphically in Figure 1 as the area under the curve to the right of the voter's retention rule.

The results of the domestic politics game can be summarized in the following proposition (with proofs of all formal results in the appendix).

**Proposition 1** *In the unique Perfect Bayesian equilibrium of the domestic politics game:*

- *if $\Psi = 0$, leaders exert effort at their ideal points, $\tilde{a}(\theta_i) = \frac{1}{\theta_i}$;*
- *congruent leaders invest greater effort into mitigation than incongruent leaders, $a^*(\underline{\theta}) > a^*(\overline{\theta})$;*
- *abatement increases in the value of office-holding, $\frac{\partial}{\partial \Psi} a^*(\theta_i) \geq 0$;*
- *congruent leaders are more likely to survive in office than incongruent leaders.*

The domestic politics game provides microfoundations for how voters can use climate policy as an informational tool to evaluate relevant aspects of leader quality. In this case, the leader's type coincides with the willingness to invest in climate mitigation, a trait corresponding with the idea that voters would prefer leaders who would take action to address climate change, all else equal (Flynn et al. 2021; Gazmararian, Mildenberger and Tingley 2023). Consequently, leaders exert effort into mitigation policy keeping in mind how voters would update their beliefs from a signal that arises from these policy choices (cf. Ashworth, Bueno de Mesquita and Friedenberg 2018; Balcazar and Kennard 2023). Voters face a classic accountability problem in which they seek to sanction leaders for pursuing agendas that would be sufficiently detrimental to the environment by exerting control at the ballot box (Ferejohn 1986), subsequently rewarding leaders who choose policies that on average bring about greater mitigation (Fearon 1999). Insofar as leaders are sensitive to the desires of this domestic audience, the value of holding office incentivizes mitigation efforts.

Additionally, the model provides an explanation for why unilateral climate action may be rational in the face of global collective action problems (Aklin and Mildenberger 2020; Kennard and Schnakenberg 2023). If domestic publics can learn about leader quality from policy implementation, the opportunity to exert mitigation efforts and signal congruence may be valuable for leaders.

## International Cooperation Game

I now introduce an international organization (IO) into the model that recommends a level of effort for leaders to pursue. The IO might represent the UNFCCC which serves as an advisory body in collecting and disseminating information about nationally determined contributions within the framework of the Paris Agreement. The goal of the IO is to reduce worldwide emissions by maximizing commitments to mitigation.

## Setup

The IO interacts with leaders by soliciting information about each leader's type $\theta_i$, which is done in the form of a report, and recommends efforts based on the reports. This can be thought of as the IO attempting to entice leaders to exert effort toward implementing (on average) better policy outcomes, conditional on the nationally determined contribution a leader submits. These reports and subsequent recommendations are then made public to all other players. As in the Paris Agreement, each leader self-reports their ability or willingness to mitigate, receives policy recommendations from an international body, but does not face punishment at the international level for noncompliance with the IO's recommendation. This organization has no enforcement power on its own, and serves a purely to aggregate and disseminate private information.

To characterize the optimal effort recommendations, I utilize the tools of mechanism design. Think of the IO as a mechanism designer tailoring its optimal policy recommendations to leaders' reports. Let leader $i$'s report be $\hat{\theta}_i \in \{\underline{\theta},\ \overline{\theta}\}$, with corresponding policy recommendation being $x(\hat{\theta}_i)$. Let $X = \sum x(\hat{\theta}_i)$. The IO designs $x(\hat{\theta}_i)$ to maximize the utilitarian objective function

$$V = \sum_i X(\hat{\theta}) - \frac{\theta_i}{2} x(\hat{\theta}_i)^2.$$

If types were perfectly known, the IO would solve the problem above by prescribing recommending

$$x(\theta_i) = \frac{n}{\theta_i},$$

which is the "first best" effort level. The first best exhibits normatively desirable properties (McAllister and Schnakenberg 2022). Recommendations are efficient, implementing a social optimum in which leaders internalize their emissions externalities, and are "variable" in the sense that each leader invests effort according to their reported willingness. Should leaders report higher (resp. lower) willingness to mitigate, the IO recommends them a more (resp. less) ambitious commitment.

However, types are private information to leaders, who must decide whether or not to reveal this type truthfully to the IO. Applying the revelation principle, I examine the class of direct mechanisms that the IO could design such that its effort recommendations would be implementable in a Perfect Bayesian equilibrium of the domestic politics game analyzed above. Thus, the IO must consider both the information and strategic constraints that leaders face. The information constraint requires that it is incentive compatible for leader $i$ to reveal its true type $\theta_i$, rather than lie and report type $\theta_i'$, given that other leaders also submit truthful reports. The strategic constraint requires that it must be weakly profitable to obey the IO's recommended level of effort.

The timing of the game is adjusted from the domestic politics game as follows:

1. Types $\theta_i$ are revealed to incumbent leaders.

2. Leaders simultaneously submit report of type $\hat{\theta}_i$ to the IO.

3. The IO recommends effort $x(\hat{\theta}_i)$ to leaders. Reports and recommendations are made public.

4. Voters observe the signal $K_i$ and bias $y_i$ and retain or replace their leaders.

5. Payoffs realized. Game ends.

A leader's strategy is now a report of its type, $\sigma_i : \{\underline{\theta}, \overline{\theta}\} \rightarrow \{\underline{\theta}, \overline{\theta}\}$, which the IO uses to determine an effort recommendation. Reports satisfy incentive compatibility and obedience constraints, to be defined formally below.

## Comments on the Model

The mechanism design approach to modeling international cooperation imposes greater structure and thus warrants further discussion of additional modeling assumptions.

**Publicity of reports.** I assume that when the IO receives self-reports from leaders, it disseminates this information worldwide. This assumption corresponds with the possibility for "naming and shaming" (e.g., Tingley and Tomz 2022; Dannenberg et al. 2023) by international and domestic audiences alike. Like the Paris Agreement, the role of the IO in this model is to provide information to leaders and voters to clarify the uncertainty around $\theta_i$. Consequently, if the mechanism is incentive compatible, voters have perfect information about leader type through the IO's reporting.

**The IO's effort recommendations.** Leaders report their type $\theta_i$, their willingness to mitigate, to the IO who then prescribes them an effort level $x(\hat{\theta}_i)$ based on this report. The report of type represents nationally determined contributions: leaders provide the IO with information about the feasibility of mitigation policy in their country based on their personal willingness to exert effort into climate policy. The IO's effort choice is analogous to $a(\theta_i)$ in the domestic politics game: the IO recommends the effort needed for the leader to implement policies to achieve the targets laid out in their nationally determined contribution.

To determine effort recommendations, the IO behaves like a utilitarian social planner, meaning that the IO hopes to realize the socially optimal reduction of emissions given what leaders report about their abilities to abate.[5] However, leaders' domestic political constraints are crucial because they dictate truthful revelation of type and obedience of the IO's recommendations.

## Analysis

I now characterize the IO's optimal effort recommendations. I begin with a benchmark case in which leaders face no electoral incentives ($\Psi = 0$), which corresponds to the case considered by McAllister and Schnakenberg (2022). Without electoral incentives to discipline their behavior, leaders' incentive constraints force the IO to "compress" its recommendations: each leader receives the same effort recommendation regardless of its reported type. I then introduce electoral incentives to show the two main results. First, climate policy cannot be "too important" electorally, or no incentive compatible agreement exists. Second, since the IO disseminates information, making leaders' reports public, electoral incentives are effectively nullified because the probability of reelection no longer depends on leaders' investments into mitigation policies. As such, the IO cannot credibly recommend that leaders invest more than their ideal points, the least ambitious investments into climate mitigation.

Without electoral incentives, the strategic tension that leaders face resembles a classic public goods problem. Declaring congruence by pledging $\hat{\theta}_i = \underline{\theta}$ and exerting greater effort into mitigation is a public good and all nations are better off if leader $i$ reduces emissions, but it is costly to leader $i$ to announce and implement such policies. Asymmetric information compounds this temptation to free ride, as leaders may have incentives to lie about their willingness to implement greater reductions. Leaders anticipate that a report about their type $\hat{\theta}_i$ maps to one of the effort levels that the IO designs, and thus have two options in choosing what to report. Leaders can either report their type $\theta_i$ truthfully, or they can mimic the other type of leader that could have been realized in their country, $\theta_i' \neq \theta_i$. Incentive compatibility requires that

$$\underbrace{x(\theta_i) + \sum_{\theta_{-i}} X_{-i}(\theta_{-i})P(\theta_{-i}) - \frac{\theta_i}{2}x(\theta_i)^2}_{\text{policy payoff from reporting } \theta_i \text{ for type } \theta_i} \geq \underbrace{x(\theta_i') + \sum_{\theta_{-i}} X_{-i}(\theta_{-i})P(\theta_{-i}) - \frac{\theta_i}{2}x(\theta_i')^2}_{\text{policy payoff from reporting } \theta_i' \text{ for type } \theta_i} \;\; \forall \theta_i \in \{\underline{\theta}, \overline{\theta}\}, \; \forall \theta_i' \in \{\overline{\theta}, \underline{\theta}\}.$$

The incentive constraint requires that the utility that leaders receive from telling the truth about their type $\theta_i$ and receiving recommendation $x(\theta_i)$ must be weakly greater than the utility from reporting type

[5] The specific form of the IO's objective function is not needed to produce the main results.

$\theta_i' \neq \theta_i$ and receiving recommendation $x(\theta_i')$, given that all other leaders report truthfully and thus in expectation their efforts are $\sum_{\theta_{-i}} X_{-i}(\theta_{-i})P(\theta_{-i})$. It is clear that incongruent leaders have no incentives to misreport and mimic congruent leaders, as doing so would allocate them a tighter commitment to mitigation than they would *ex ante* prefer. However, congruent leaders may benefit from misreporting in order to avoid greater costs associated with more committed mitigation efforts, thus mimicking incongruent leaders. Consequently, every leader would be tempted to report $\hat{\theta}_i = \overline{\theta}$; the IO fails to screen congruent leaders from incongruent leaders. To satisfy incentive compatibility constraints, the IO must prescribe a "compressed" effort recommendation to each leader regardless of their reported willingness to mitigate. Put differently, the IO's recommendations are type-invariant within countries (cf. Harrison and Lagunoff 2017; McAllister and Schnakenberg 2022).[6]

**Lemma 1** *Suppose* $\Psi = 0$. *Any incentive compatible mechanism is "compressed:"* $x(\underline{\theta}) = x(\overline{\theta})$. *The recommendation takes the form*

$$x^*(\hat{\theta}_i) = \frac{n}{q\underline{\theta} + (1-q)\overline{\theta}}.$$

Counterintuitively, congruent leaders obstruct efficient international cooperation. When designing incentive compatible effort recommendations, the IO has to worry about climate-friendly leaders not willing to contribute their share to abating the effects of climate change, not the climate laggards.

I now introduce electoral incentives, $\Psi > 0$. A key jump from the domestic politics game into the international cooperation game is that leader reports and subsequent effort recommendations are public to voters. Moreover, if the agreement is incentive compatible such that it induces truthful revelation of leader type, voters learn the types of their leaders with certainty through these reports, The voter's posterior belief $\mu$ about leader congruence is either 1 or zero. Subsequently, whether leader $i$ survives in office does not depend on policy implementation $K_i$ but rather depends entirely on the realization of the valence shock: the voter retains the leader if $y_i \geq q - \mu$ which occurs with probability $\frac{\gamma - q + \mu}{2\gamma}$. Information constraints for

[6] Because countries are symmetric (the values of $\underline{\theta}$, $\overline{\theta}$, and $q$ are constant across countries), "within-country compression" and "across-country compression" are indistinguishable in the model but are conceptually distinct. Across-country compression means that every country receives the same effort recommendation. This behavior is reminiscent of the broader-deeper tradeoff (Downs, Rocke and Barsoom 1998) or the law of the least ambitious program (Hovi, Ward and Grundig 2015). If countries were not symmetric, across-country compression would not hold, but within-country compression would.

congruent an incongruent leaders respectively now reflect this:

$$\underbrace{x(\underline{\theta}) + \sum_{\theta_{-i}} X_{-i}(\theta_{-i})P(\theta_{-i}) - \frac{\underline{\theta}}{2}x(\underline{\theta})^2}_{\text{policy payoff from truthfully reporting } \theta_i=\underline{\theta}} + \underbrace{\frac{1}{2\gamma}\Psi}_{\substack{\text{net electoral benefit for} \\ \text{reporting } \theta_i=\underline{\theta}}} \geq \underbrace{x(\overline{\theta}) + \sum_{\theta_{-i}} X_{-i}(\theta_{-i})P(\theta_{-i}) - \frac{\underline{\theta}}{2}x(\overline{\theta})^2}_{\text{policy payoff from reporting } \theta_i=\overline{\theta}}.$$

$$\underbrace{x(\overline{\theta}) + \sum_{\theta_{-i}} X_{-i}(\theta_{-i})P(\theta_{-i}) - \frac{\overline{\theta}}{2}x(\overline{\theta})^2}_{\text{policy payoff from truthfully reporting } \theta_i=\overline{\theta}} \geq \underbrace{x(\underline{\theta}) + \sum_{\theta_{-i}} X_{-i}(\theta_{-i})P(\theta_{-i}) - \frac{\overline{\theta}}{2}x(\underline{\theta})^2}_{\text{policy payoff from reporting } \theta_i=\underline{\theta}} + \underbrace{\frac{1}{2\gamma}\Psi}_{\substack{\text{net electoral benefit for} \\ \text{reporting } \theta_i=\underline{\theta}}}.$$

Leader behavior differs from the benchmark where $\Psi = 0$ because leaders' actions are now relevant within the context of their interaction with their voters. The reports that they submit to the IO have downstream domestic consequences, namely that the IO resolves the uncertainty regarding each leader's type. Any leader who reports to be congruent would be reelected with probability $\frac{\gamma-q+1}{2\gamma}$; any leader who reports to be incongruent would be reelected with probability $\frac{\gamma-q}{2\gamma}$. In other words, the publicity of reports renders the policy outcome $K_i$ irrelevant for the leader's reelection prospects because voters have access to selection-relevant information through the IO.

It is immediately clear that compressed recommendations are never incentive compatible for incongruent leaders, because they could always do better by claiming to be congruent and thus increase their chances of reelection. Therefore, it must be the case that $x^*(\underline{\theta}) > x^*(\overline{\theta})$, or that congruent leaders must always invest more effort than incongruent leaders. Since incongruent leaders win reelection with a lower probability, they must be compensated through a more lenient effort recommendation. Conversely, because congruent types benefit electorally from truthful revelation, they must also shoulder greater costs of abatement by exerting more effort into mitigation policy.

Introducing electoral incentives complicates intuition about the conditions under which leaders would report truthfully. Congruent leaders may mimic incongruent leaders to incur fewer mitigation costs, but incongruent leaders may mimic congruent leaders to increase their reelection chances. It is thus not clear *ex ante* as to which type's incentive constraint would bind. Instead, a range on $\Psi$ can be found such that both constraints would be simultaneously satisfied. This range is

$$\gamma(x(\underline{\theta}) - x(\overline{\theta}))(\underline{\theta}(x(\underline{\theta}) + x(\overline{\theta})) - 2) \leq \Psi \leq \gamma(x(\underline{\theta}) - x(\overline{\theta}))(\overline{\theta}(x(\underline{\theta}) + x(\overline{\theta})) - 2).$$

This range defines the extent to which leaders value electoral benefits relative to utility from exerting effort on climate policy. The lower bound on $\Psi$ comes from the congruent leader's constraint. By lying, a congruent leader incurs fewer costs from exerting effort into mitigation policy, but misreporting comes at the expense of decreased electoral odds. If electoral incentives are large enough, then the congruent leader finds it sufficient to forego the utility from a more lenient effort recommendation – thus reversing her strategic logic from the case without electoral incentives – and report her true type. Conversely, incentive compatibility for the incongruent leader requires that the value of domestic office is not too large, establishing the upper bound. If electoral incentives are hefty, incongruent leaders would be incentivized to mimic congruent leaders, even if it means shouldering greater costs of mitigation than they would prefer, in order to increase the chances of remaining in power.

Evidently, what can be supported as an incentive compatible effort level depends on how leaders trade off mitigating the effects of climate change and maintaining office. To the extent that it can, the IO seeks to design efforts $x(\underline{\theta})$ and $x(\overline{\theta})$ to maximize the range of electoral incentives such that truthful revelation is incentive compatible, knowing the domestic consequences. For example, the first best could be sustained as an incentive compatible recommendation if $\frac{n\gamma(\overline{\theta}-\underline{\theta})(n\underline{\theta}+\overline{\theta}(n-2))}{\overline{\theta}^2\underline{\theta}} \leq \Psi \leq \frac{n\gamma(\overline{\theta}-\underline{\theta})(n\overline{\theta}+\underline{\theta}(n-2))}{\overline{\theta}\underline{\theta}^2}$. However, since leaders cannot exert infinite amounts of effort into mitigation, the IO's tools in optimizing this tradeoff are limited. This is particularly difficult when thinking about cases in which leaders care primarily about domestic political survival, or where $\Psi \to \infty$. The first result is that, if survival concerns become too important, then no incentive compatible mechanism exists.

**Lemma 2** *No incentive compatible mechanism exists if* $\Psi > \frac{\gamma(\overline{\theta}\omega-1)^2}{\overline{\theta}}$.

Lemma 2 establishes the limit on the tradeoff between investing effort into mitigation and holding office. Recall that incongruent leaders find it incentive compatible to report truthfully if $\Psi \leq \gamma(x(\underline{\theta})-x(\overline{\theta}))(\overline{\theta}(x(\underline{\theta})+x(\overline{\theta}))-2)$, because reporting congruence is electorally valuable. Since leaders can exert a maximum effort level $\omega$, this constraint is bounded. Specifically, it is maximized if $x(\underline{\theta}) = \omega$ and $x(\overline{\theta}) = \frac{1}{\overline{\theta}}$. In this mechanism, the IO recommends congruent types to exert the maximum possible effort, while incongruent types enact their ideal point. This establishes a largest possible value for $\Psi$ such that incongruent leaders would find it in their interest to truthfully report their type; if $\Psi$ is larger than this upper bound, it is never politically valuable to tell the truth. Incongruent leaders would rather pretend to be congruent in order to increase their electoral odds, violating incentive compatibility.[7]

[7] What if we were to eschew the requirement that leaders report truthfully? Suppose $\Psi > \frac{\gamma(\overline{\theta}\omega-1)^2}{\overline{\theta}}$, so leaders report to the IO that they are congruent, regardless of their true type. It is therefore optimal for the IO to recommend the largest mitigation investment that could be supported by some type of pooling: $x^*(\hat{\underline{\theta}}) = \frac{n}{q\underline{\theta}+(1-q)\overline{\theta}}$, as in Lemma 1. Since all leaders report $\hat{\theta}_i = \underline{\theta}$, the IO provides all leaders with the same recommendation, accounting for the fact that there is some chance that it may be facing incongruent leaders mimicking congruent leaders. Moreover, because incentive compatibility no longer

So far, I have only examined whether leaders find it incentive compatible to reveal their true type to the IO. For pledges to be implementable, I must also consider leaders' incentives to follow the IO's effort recommendations. I examine *ex post* obedience constraints: conditional on the public reports of type, leaders must be willing to implement their recommended policies rather than some other profitable deviation.[8] The obedience constraint thus requires that

$$\underbrace{x(\theta_i) + \sum_{\theta_{-i}} X_{-i}(\theta_{-i}) - \frac{\theta_i}{2} x(\theta_i)^2}_{\text{policy payoff from obeying IO as type } \theta_i} + \underbrace{\frac{\gamma - q + \mu}{2\gamma}\Psi}_{\substack{\text{electoral payoff for} \\ \text{obeying IO as type } \theta_i}} \geq \max_d \underbrace{d + \sum_{\theta_{-i}} X_{-i}(\theta_{-i}) - \frac{\theta_i}{2} d^2}_{\text{policy payoff from maximal deviation for type } \theta_i} + \underbrace{\frac{\gamma - q + \mu}{2\gamma}\Psi}_{\substack{\text{electoral payoff for} \\ \text{deviating as type } \theta_i}}.$$

The publicity of leader reports nullifies electoral incentives for all leaders *ex post*, regardless of their type. That is, because the IO announces leaders' reports and its recommendations, resolving uncertainty around type, leaders' electoral odds are no longer connected to the effort that they exert into mitigation policy. This therefore also removes any incentives for leaders to comply with particularly ambitious investments into mitigation because the IO washes out the leader's accountability relationship with the voter. The only effort levels that satisfy leaders' obedience constraints are their ideal points, $\tilde{a}(\theta_i) = \frac{1}{\theta_i}$. This yields the second result, which is that the revelation of information deactivates the electoral mechanism through which leaders exert effort in the hopes of increasing their electoral odds.

**Lemma 3** *Recommendations $x^*(\theta_i)$ satisfy leaders' obedience constraints if and only if $x^*(\theta_i) = \frac{1}{\theta_i} = \tilde{a}(\theta_i)$.*

The result follows directly from the fact that the IO reveals electorally relevant information to voters by making information related to leader type publicly available. The IO breaks the accountability chain between leaders and voters, so leaders have no incentives to make more ambitious commitments than their ideal efforts. Consequently, any recommendation other than the ideal effort level would not be individually rational for leaders to obey, pinning down the optimal effort recommendations. Counterintuitively, leaders would always pursue greater investments into climate mitigation policy on their own rather than within the framework of an international agreement because of the presence of electoral incentives.

**Proposition 2** *In an equilibrium implemented by the international cooperation game:*

- *leaders exert effort at their ideal points, $x^*(\theta_i) = \tilde{a}(\theta_i) = \frac{1}{\theta_i}$, only if $\Psi \in [0, \frac{\gamma(\overline{\theta}-\underline{\theta})^2}{\overline{\theta}\underline{\theta}^2}]$;*

holds for incongruent types and the IO has no way to screen them from congruent types, the voter learns nothing from the IO's announcements of types. Consequently, there is no updating, $\mu = q$, and leaders survive in office with probability $\frac{1}{2}$.

[8] This implies perfect compliance with the IO's choice of effort for leader $i$, rather than pursuing some other profitable deviation. Obedience also rules out any chance of exiting from the agreement. If a leader's obedience constraint is satisfied, she has no incentives to defect on the agreement after the IO has made a policy recommendation, conditional on knowing her type and conditional on all other countries obeying the IO's recommendations.

- *congruent leaders invest greater effort into mitigation than incongruent leaders,* $x^*(\underline{\theta}) > x^*(\overline{\theta})$*;*
- *the value of office-holding has no effect on mitigation efforts,* $\frac{\partial}{\partial \Psi} x^*(\theta_i) = 0$*;*
- *congruent leaders are more likely to survive in office than incongruent leaders;*
- *expected aggregate efforts are less than those in the domestic politics game,* $E[X^*] \leq E[A^*]$*.*

Combining insights from leaders' incentive compatibility and obedience constraints determines the IO's optimal effort recommendations. Lemma 3 identifies that, because the IO publicizes information about leader type and subsequently breaks the accountability relationship that would have incentivized increased mitigation efforts, the only recommendations that would satisfy the obedience constraint are leaders' ideal points, $x^*(\theta_i) = \tilde{a}(\theta_i) = \frac{1}{\theta_i}$. Moreover, while this recommendation is always incentive compatible for congruent leaders, the ideal effort level is only incentive compatible for incongruent leaders if $\Psi \leq \frac{\gamma(\overline{\theta}-\underline{\theta})^2}{\overline{\theta}\underline{\theta}^2}$. Importantly, the features of the institution counteract the incentives for more ambitious climate action. The revelation of electorally relevant information, leaders' willingness to mitigate, means that the IO resolves the voter's domestic selection problem. However, this impacts how leaders interface with the institution in the first place: specifically, public information about leader type in the form of nationally determined contributions voids leaders' electoral incentives to exert effort, leaving them with little reason to comply with ambitious policies *ex post*.

Figure 2 summarizes the key strategic implications across both the domestic politics game and the international cooperation game. The solid black characterize leaders' ideal efforts, the equilibrium of the domestic politics game when $\Psi = 0$. Introducing these incentives increases leaders' mitigation efforts because they now have the ability to signal congruence to a domestic audience. This is illustrated by the solid purple lines in which optimal effort $a^*(\theta_i)$ in the domestic politics game is increasing in the value of $\Psi$.

In the international cooperation game, the IO strives to maximize international commitments to mitigation policy. The normative ideal would be for each leader to exert effort at the social optimum, which is captured by the purple crosses in Figure 2. Contrasted with leaders' ideal points, leaders internalize their emissions externalities in the first best. However, institutions like Paris rely on leaders to self-report their willingness to mitigate, which constrains the ability for the IO to implement the social optimum. Indeed, since leaders have private information about their willingness to mitigate, the IO must incentivize truthful revelation of this information.

If there are no electoral incentives, introducing informational constraints compresses the IO's recommended effort levels to the black triangle. Intuitively, the compressed recommendation is a weighted average

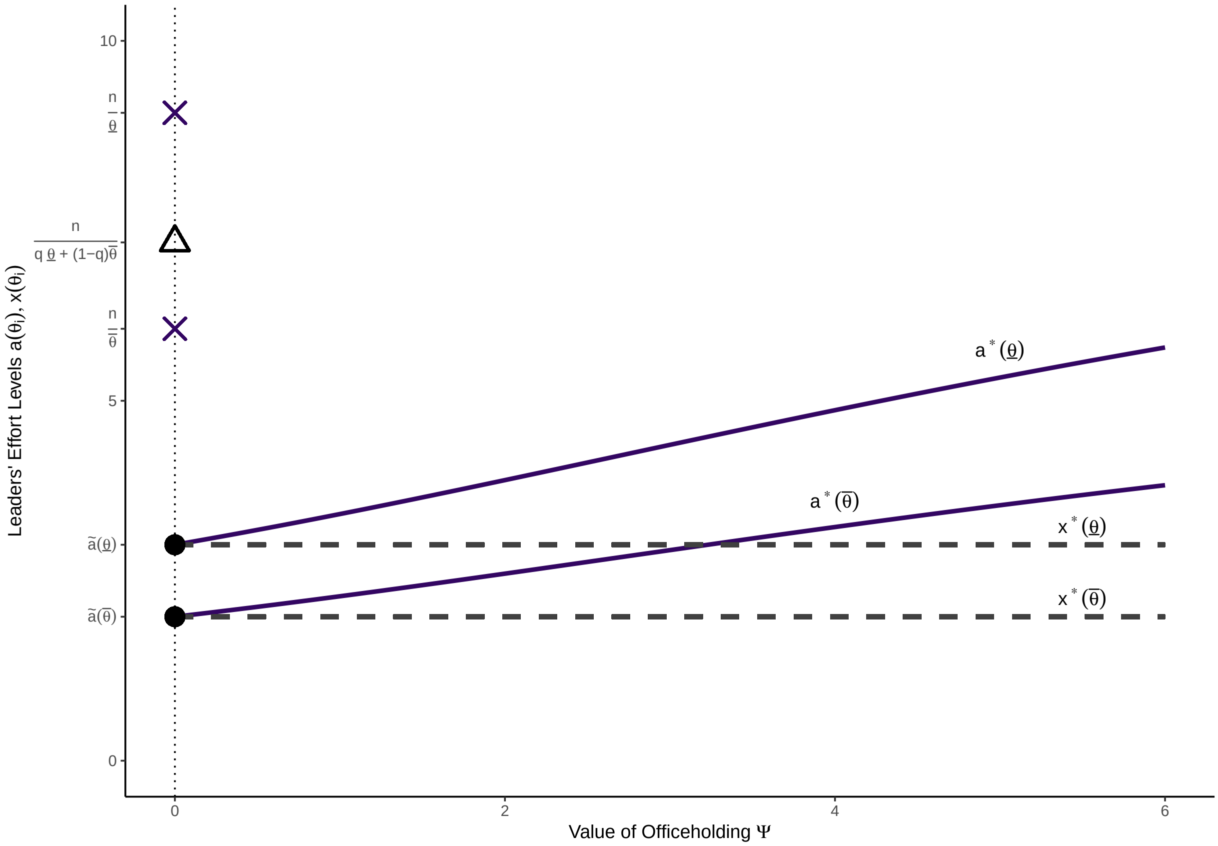


Figure 2: Optimal Efforts in the Domestic Politics and International Cooperation Games

The solid black points are leaders' ideal efforts, $\tilde{a}(\theta_i) = \frac{1}{\theta_i}$. The solid purple lines are leaders' equilibrium efforts in the domestic politics game as a function of the value of holding office, $a^*(\theta_i)$. The purple crosses are leaders first best efforts in the international cooperation game, $\frac{n}{\theta_i}$. The black triangle is the IO's compressed recommendation in the international cooperation game, $\frac{n}{q\underline{\theta}+(1-q)\overline{\theta}}$. The dashed grey lines are the IO's optimal recommendations in the domestic politics game, $x^*(\theta_i) = \frac{1}{\theta_i}$.

of each leader-type's investment into mitigation policy in a first best world; however, the compressed recommendation fails to be optimal for two reasons. First, the IO can only recommend a compressed effort level if leaders do not care about reelection. If $\Psi > 0$, incongruent leaders do not find it incentive compatible to truthfully report their willingness to mitigate. Second, leaders face strategic constraints on top of their informational constraints. Each leader must individually be willing to exert the recommended effort, and the compressed recommendation fails this obedience criterion. The equilibrium of the international cooperation game is therefore illustrated by the dashed grey lines, in which leaders simply implement their ideal points. The IO fails to motivate leaders to exert more effort than they would in a canonical public goods problem, and worsens mitigation efforts relative to what they would be in the domestic politics game in which leaders have electoral incentives to behave more ambitiously.

## Who Joins Paris?

The Paris Agreement has nearly universal membership across the global stage. This analysis would suggest that, if leaders were motivated to maximize their commitments to mitigation, they would be better off without international cooperation that facilitates the revelation of private information. What then incentivizes leaders to join such an agreement? I now consider the *ex ante* decision for leaders to either join the agreement, playing the international cooperation game, or not, playing the domestic politics game. I perform this calculation prior to leaders learning their type, so that I abstract away from the potential signaling encoded in a leader's decision to join the agreement in the first place. If leader $i$ joins, she receives

$$EU_i(\text{join}) = q\Big[\frac{1}{\underline{\theta}} + E[X^*_{-i}] - \frac{1}{2\underline{\theta}} + \frac{\gamma - q + 1}{2\gamma}\Psi\Big] + (1-q)\Big[\frac{1}{\overline{\theta}} + E[X^*_{-i}] - \frac{1}{2\overline{\theta}} + \frac{\gamma - q}{2\gamma}\Psi\Big].$$

By contrast, remaining out of the agreement and playing the domestic politics game yields her

$$\begin{aligned} EU_i(\neg\text{join}) = q\Big[a^*(\underline{\theta}) + E[A^*_{-i}] - \frac{\underline{\theta}}{2}a^*(\underline{\theta})^2 + \frac{\gamma - q + \int_{-1+q}^{q} \Phi(\sqrt{\beta}(a^*(\underline{\theta}) - \hat{K}^*(y_i)))\ dy}{2\gamma}\Psi\Big] \\ + (1-q)\Big[a^*(\overline{\theta}) + E[A^*_{-i}] - \frac{\overline{\theta}}{2}a^*(\overline{\theta})^2 + \frac{\gamma - q + \int_{-1+q}^{q} \Phi(\sqrt{\beta}(a^*(\overline{\theta}) - \hat{K}^*(y_i)))\ dy}{2\gamma}\Psi\Big]. \end{aligned}$$

Clearly, the incentives to join the agreement depend on how leaders benefit electorally from information revelation. In particular, the decision to join boils down to two parameters: the chance that the leader will be competent $q$, and the salience of climate policy relative to other electorally important issues $\gamma$. The next result formalizes this relationship.

**Proposition 3** *If leaders are expected to be incongruent $q < \frac{1}{2}$, there exists a threshold $\overline{\gamma}$ such that leaders join the agreement whenever $\gamma > \overline{\gamma}$.*

Proposition 3 states that leaders prefer to join when they expect to be incongruent and climate policy is not too salient. Leaders who expect to be incongruent are only willing to join the agreement when climate policy is not salient because it means that these leaders could still win the election based on their popularity on other electorally relevant issues. Recall that the IO resolves the voter's selection problem, which detracts from incongruent leaders' electoral odds relative to what would happen in the domestic politics game. Indeed, when party to the agreement, incongruent leaders are reelected with probability $\frac{\gamma-q}{2\gamma} < \frac{\gamma-q+\int_{-1+q}^{q} \Phi(\sqrt{\beta}(a^*(\bar{\theta})-\hat{K}^*(y_i)))\ dy}{2\gamma}$, their probability of survival in the domestic politics game. Moreover, leaders are expected to invest less in mitigation when party to the agreement, so they incur fewer costs of exerting effort. Therefore, when mitigation efforts become less salient to voters relative to other issues, captured by increasing $\gamma$, it means that the leader could remain in office for reasons external to climate policy, so the value of the agreement increases for incongruent leaders.

If leaders are expected to be congruent, the effects of joining are less clear to parse because there is electoral value from joining but there are policy benefits from playing the domestic politics game. Congruent leaders benefit from the agreement because it enhances their electoral odds: by resolving the voter's uncertainty about leader type, the IO ensures that any leader who reports to be congruent is more likely to survive in office. This means that there are positive electoral benefits for congruent leaders for joining the agreement. However, leaders may also wish not to cooperate, playing the domestic politics game instead, when climate policy becomes less salient. Indeed, when $\gamma$ increases, leaders invest less effort into mitigation policy because the chance that climate policy affects the voter's election decision diminishes. As mitigation becomes a less salient issue, the relative electoral benefits that congruent leaders reap decrease.

This result can help to rationalize the broad membership that the Paris Agreement enjoys. While the salience of climate change has increased over time (Egan, Konisky and Mullin 2022), the demand for climate policy and its subsequent effect on electoral outcomes continue to be small (Bernauer and Gampfer 2015; Egan and Mullin 2017), implying lower salience relative to other electorally motivating issues. If $\gamma$ is large, even those unwilling to mitigate may find the stakes of joining the agreement low. Indeed, such leaders benefit because they can exert less effort into mitigation relative to what they would do in an equilibrium without the agreement, and they can salvage their electoral odds through their popularity on other issues. Moreover, large $\gamma$ also implies a weakening of the accountability mechanism between leaders and their domestic publics, as policy outcomes through this channel are not as politically dispositive. This may clarify why leaders in places with weaker accountability relationships, for example in less democratic societies, were willing to join the Paris Agreement.

## Discussion and Conclusion

By way of conclusion, I consider the implications of the institutional design of the agreement analyzed in this paper for contemporary climate governance.

One hope for the Paris Agreement was that its transparency would facilitate deeper cooperation because climate laggards would be "named and shamed," incurring reputational costs for noncompliance (Falkner 2016; Dannenberg et al. 2023). However, beyond conjecturing the existence of these reputational costs, scholars have not unpacked their strategic microfoundations. Experimental work demonstrates that the public and climate elites support naming shaming as a tool (Tingley and Tomz 2022; Dannenberg et al. 2023), but where would such naming and shaming come from, and how would the price that leaders pay manifest? A flexible interpretation of the model suggests a domestic microfoundation in which voters shame leaders who may be perceived as incongruent with the threat of electoral removal. However, contrary to conventional wisdom, I demonstrate that leaders invest *less* effort into climate mitigation when an international organization provides relevant information that could facilitate naming and shaming.

Since the beginning of the pledge-and-review process, experts have assessed that nations have failed to "deliver on their promises" to meet the Paris goal of limiting warming to below 1.5-2°C, calling for more ambitious climate action (UNEP 2021; UNFCCC 2023). These reports evaluate progress on observed policy outputs, stochastic in the model; the agreement itself aspires to motivate more ambitious policy inputs, greater effort investments. Equilibrium behavior in the model is consistent with the idea that, on average, observed policy outcomes fall short of countries' achievement of Paris goals because the effort they put into meeting those goals is minimal. The model demonstrates that more ambitious mitigation efforts may not have been credible by design.

This study has scrutinized the institutional features central to Paris as the newest addition to the "regime complex for climate change" (Keohane and Victor 2011) and compared equilibrium efforts in this setting to one without international cooperation. Two fruitful avenues for further research emerge. The first concerns the comparative analysis of alternative climate governance institutions (e.g., Nordhaus 2015; Harstad 2023); this paper suggests that voluntary commitments to mitigation efforts and publicity of nationally determined contributions do not further progress toward maximizing mitigation commitments. However there are certainly other institutional features that may be relevant in determining mitigation efforts, either empirical (e.g., Kyoto) or theoretical. These institutions may have other benefits or costs to climate cooperation that may alter the relative attractiveness of the Paris framework.

Second, scholars should continue to interrogate how domestic accountability affects international cooperation. The model relies on an accountability mechanism through which a voter seeks to select congruent leaders and incentivize mitigation efforts. These types of accountability channels are often associated with democracies (Przeworski, Manin and Stokes 1999). While the model can accommodate slack in the accountability relationship, for example by taking $\beta \to 0$ so that policy outcomes are completely uninformative of leader type, future work could explore additional mechanisms in weaker accountability contexts and characterize how those domestic institutions affect international cooperation.

## References


Aklin, Michaël and Matto Mildenberger. 2020. "Prisoners of the Wrong Dilemma: Why Distributive Conflict, Not Collective Action, Characterizes the Politics of Climate Change." *Global Environmental Politics* 20(4):4–27.

Ansolabehere, Stephen and David M. Konisky. 2014. *Cheap and Clean: How Americans Think about Energy in the Age of Global Warming.* Cambridge: MIT Press.

Ashworth, Scott and Ethan Bueno de Mesquita. 2006. "Monotone Comparative Statics for Models of Politics." *American Journal of Political Science* 50(1):214–231.

Ashworth, Scott, Ethan Bueno de Mesquita and Amanda Friedenberg. 2018. "Learning about Voter Rationality." *American Journal of Political Science* 62(1):37–54.

Balcazar, Carlos Felipe and Amanda Kennard. 2023. "Climate Change and Political Mobilization: Theory and Evidence from India.".

Barrett, Scott. 2003. *Environment and Statecraft: The Strategy of Environmental Treaty-Making.* Oxford: Oxford University Press.

Bechtel, Michael M. and Kenneth F. Scheve. 2013. "Mass support for global climate agreements depends on institutional design." *Proceedings of the National Academy of Sciences* 110(34):13763–13768.

Bernauer, Thomas and Robert Gampfer. 2015. "How robust is public support for unilateral climate policy?" *Environmental Science & Policy* 54:316–330.

Besley, Timothy. 2006. *Principled Agents?: The Political Economy of Good Government.* Oxford: Oxford University Press.

Besley, Timothy and Robin Burgess. 2002. "The Political Economy of Government Responsiveness: Theory and Evidence from India." *The Quarterly Journal of Economics* 117(4):1415–1451.

Bisbee, James H., James R. Hollyer, B. Peter Rosendorff and James Raymond Vreeland. 2019. "The Millennium Development Goals and Education: Accountability and Substitution in Global Assessment." *International Organization* 73(3):547–578.

Bodansky, Daniel. 2016. "The Legal Character of the Paris Agreement." *Review of European, Comparative & International Environmental Law* 25(2):142–150.

Breetz, Hanna, Matto Mildenberger and Leah Stokes. 2018. "The political logics of clean energy transitions." *Business and Politics* 20(4):492–522.

Dai, Xinyuan. 2005. "Why Comply? The Domestic Constituency Mechanism." *International Organization* 59(2):363–398.

Dannenberg, Astrid, Marcel Lumkowsky, Emily K. Carlton and David G. Victor. 2023. "Naming and shaming as a strategy for enforcing the Paris Agreement: The role of political institutions and public concern." *Proceedings of the National Academy of Sciences* 120(40):e2305075120.

Depledge, Joanna. 2006. "The Opposite of Learning: Ossification in the Climate Change Regime." *Global Environmental Politics* 6(1):1–22.

Downs, George W., David M. Rocke and Peter N. Barsoom. 1998. "Managing the Evolution of Multilateralism." *International Organization* 52(2):397–419.

Egan, Patrick J, David M Konisky and Megan Mullin. 2022. "Ascendant Public Opinion: The Rising Influence of Climate Change on Americans' Attitudes about the Environment." *Public Opinion Quarterly* 86(1):134–148.

Egan, Patrick J. and Megan Mullin. 2017. "Climate change: US public opinion." *Annual Review of Political Science* 20:209–227.

Falkner, Robert. 2016. "The Paris Agreement and the new logic of international climate politics." *International Affairs* 92(5):1107–1125.

Fearon, James D. 1999. Electoral Accountability and the Control of Politicians: Selecting Good Types versus Sanctioning Poor Performance. In *Democracy, Accountability, and Representation*, ed. Adam Przeworski, Bernard Manin and Susan C. Stokes. Cambridge Studies in the Theory of Democracy Cambridge: Cambridge University Press pp. 55–97.

Ferejohn, John. 1986. "Incumbent Performance and Electoral Control." *Public Choice* 50(1/3):5–25.

Flynn, Cassie, Eri Yamasumi, Stephen Fisher, Dan Snow, Zack Grant, Martha Kirby, Peter Browning, Moritz Rommerskirchen and Inigo Russell. 2021. Peoples' Climate Vote. Technical report UNDP and University of Oxford.

Fox, Justin and Richard Van Weelden. 2012. "Costly transparency." *Journal of Public Economics* 96(1):142–150.

Gaikwad, Nikhar, Federica Genovese and Dustin Tingley. 2022. "Creating Climate Coalitions: Mass Preferences for Compensating Vulnerability in the World's Two Largest Democracies." *American Political Science Review* 116(4):1165–1183.

Gailmard, Sean and John W. Patty. 2019. "Preventing Prevention." *American Journal of Political Science* 63(2):342–352.

Gazmararian, Alexander F. and Dustin Tingley. 2023. *Uncertain Futures: How to Unlock the Climate Impasse*. New York: Cambridge University Press.

Gazmararian, Alexander F., Matto Mildenberger and Dustin Tingley. 2023. "Public Opinion Foundations of the Clean Energy Transition.".

Harrison, Rodrigo and Roger Lagunoff. 2017. "Dynamic Mechanism Design for a Global Commons." *International Economic Review* 58(3):751–782.

Harstad, Bård. 2023. "Pledge-and-review bargaining." *Journal of Economic Theory* 207(C).

Hoffmann, Matthew J. 2008. The Global Regime: Current Status of and Quo Vadis for Kyoto. In *A globally integrated climate policy for Canada*, ed. Steven F. Bernstein, Jutta Brunnee, David G. Duff and Andrew J. Greene. Toronto: University of Toronto Press pp. 137–157.

Hollyer, James R. and B. Peter Rosendorff. 2011. "Why Do Authoritarian Regimes Sign the Convention Against Torture? Signaling, Domestic Politics and Non-Compliance." *Quarterly Journal of Political Science* 6(3–4):275–327.

Hollyer, James R., B. Peter Rosendorff and James Raymond Vreeland. 2015. "Transparency, Protest, and Autocratic Instability." *American Political Science Review* 109(4):764–784.

Hovi, Jon, Hugh Ward and Frank Grundig. 2015. "Hope or Despair? Formal Models of Climate Cooperation." *Environmental & Resource Economics* 62(4):665–688.

Jacquet, Jennifer and Dale Jamieson. 2016. "Soft but significant power in the Paris Agreement." *Nature Climate Change* 6(7):643–646.

Kelley, Judith and Beth Simmons. 2015. "Politics by Number: Indicators as Social Pressure in International Relations." *American Journal of Political Science* 59(1):55–70.

Kelley, Judith G. and Beth A. Simmons. 2019. "Introduction: The Power of Global Performance Indicators." *International Organization* 73(3):491–510.

Kennard, Amanda and Keith E. Schnakenberg. 2023. "Comment: Global Climate Policy and Collective Action." *Global Environmental Politics* 23(1):133–144.

Keohane, Robert O. 1984. *After Hegemony: Cooperation and Discord in the World Political Economy.* Princeton: Princeton University Press.

Keohane, Robert O. and David G. Victor. 2011. "The Regime Complex for Climate Change." *Perspectives on Politics* 9(1):7–23.

Keohane, Robert O. and David G. Victor. 2016. "Cooperation and discord in global climate policy." *Nature Climate Change* 6:570–575.

Keohane, Robert O. and Michael Oppenheimer. 2016. "Paris: Beyond the Climate Dead End through Pledge and Review?" *Politics and Governance* 4(3):142–151.

McAllister, Jordan H. and Keith E. Schnakenberg. 2022. "Designing the Optimal International Climate Agreement with Variability in Commitments." *International Organization* 76(2):469–486.

Mildenberger, Matto. 2020. *Carbon Captured: How Business and Labor Control Climate Politics.* Cambridge: The MIT Press.

Nordhaus, William. 2015. "Climate Clubs: Overcoming Free-Riding in International Climate Policy." *American Economic Review* 105(4):1339–1370.

Prat, Andrea. 2005. "The Wrong Kind of Transparency." *The American Economic Review* 95(3):862–877.

Przeworski, Adam, Bernard Manin and Susan C. Stokes, eds. 1999. *Democracy, Accountability, and Representation.* Cambridge Studies in the Theory of Democracy Cambridge: Cambridge University Press.

Slechten, Aurélie. 2020. "Environmental Agreements under Asymmetric Information." *Journal of the Association of Environmental and Resource Economists* 7(3):455–481.

Stern, Nicholas. 2007. *The Economics of Climate Change: The Stern Review.* Cambridge: Cambridge University Press.

Stokes, Leah C. 2016. "Electoral Backlash against Climate Policy: A Natural Experiment on Retrospective Voting and Local Resistance to Public Policy." *American Journal of Political Science* 60(4):958–974.

Stokes, Leah C. 2020. *Short Circuiting Policy: Interest Groups and the Battle Over Clean Energy and Climate Policy in the American States*. New York: Oxford University Press.

Tingley, Dustin and Michael Tomz. 2022. "The Effects of Naming and Shaming on Public Support for Compliance with International Agreements: An Experimental Analysis of the Paris Agreement." *International Organization* 76:445–468.

UNEP. 2021. Emissions Gap Report 2021: The Heat is On - A World of Climate Promises Not Yet Delivered. Technical report UNEP Copenhagen Climate Centre Nairobi: .

UNFCCC. 2023. Technical dialogue of the first global stocktake. Technical report United Arab Emirates: .

Victor, David G. 2011. *Global Warming Gridlock: Creating More Effective Strategies for Protecting the Planet*. Cambridge: Cambridge University Press.

von Stein, Jana. 2008. "The International Law and Politics of Climate Change: Ratification of the United Nations Framework Convention and the Kyoto Protocol." *Journal of Conflict Resolution* 52(2):243–268.

# Appendix

**Proposition 1** *In the unique Perfect Bayesian equilibrium of the domestic politics game:*

- *if $\Psi = 0$, leaders invest effort at their ideal points, $\tilde{a}(\theta_i) = \frac{1}{\theta_i}$;*
- *congruent leaders invest greater effort into mitigation than incongruent leaders, $a^*(\underline{\theta}) > a^*(\overline{\theta})$;*
- *abatement increases in the value of office-holding, $\frac{\partial}{\partial \Psi} a^*(\theta_i) \geq 0$;*
- *congruent leaders are more likely to survive in office than incongruent leaders.*

We will prove Proposition 1 with a series of claims.

**Claim 1** *The unique equilibrium of the domestic politics game is characterized by a double $(\underline{a}, \overline{a})$, which represents leader-type $\theta_i$'s policy choices that forms a Bayesian Nash equilibrium given the policy choices of leader-types in other countries $\theta_{-i}$. The voter in country $i$ retains the leader if and only if $K_i \geq \hat{K}(y_i)$.*

**Proof of Claim 1:** Begin by noting that because countries are symmetric, all leaders with type $\underline{\theta}$ will choose the same policy, as will all leaders with type $\overline{\theta}$. It suffices to omit the subscript $i$ because, conditional on type, leaders behave identically.

Voter $i$ adopts a decision rule in which he retains the leader if and only if

$$P(\underline{\theta}|K_i) + y_i \geq q.$$

The voters in each country need to have conjectures about the policies chosen by each leader-type. Denote these by $(\hat{a}(\underline{\theta}), \hat{a}(\overline{\theta}))$. Posterior beliefs about leader $i$'s type given the observed value of the signal are

$$P(\underline{\theta}|K_i) = \frac{q\phi(\sqrt{\beta}(K_i - \hat{a}(\underline{\theta})))}{q\phi(\sqrt{\beta}(K_i - \hat{a}(\underline{\theta}))) + (1-q)\phi(\sqrt{\beta}(K_i - \hat{a}(\overline{\theta}))}.$$

Conditional on some value of the valence shock $y_i$, the voter is thus exactly indifferent between retaining the incumbent leader and replacing her when

$$K(y_i) = \frac{\hat{a}(\underline{\theta}) + \hat{a}(\overline{\theta})}{2} + \frac{\log\left(\frac{(1-q)(q-y_i)}{q(1-q+y_i)}\right)}{\beta(\hat{a}(\underline{\theta}) - \hat{a}(\overline{\theta}))}.$$

The likelihood ratio $\frac{\phi(\sqrt{\beta}(K_i - \hat{a}(\underline{\theta})))}{\phi(\sqrt{\beta}(K_i - \hat{a}(\overline{\theta}))}$ is increasing in the signal $K_i$. Therefore, the voter in country $i$ retains his leader if and only if $K_i \geq \hat{K}$. Also note that if $y_i > q$ then $K(y_i) \to -\infty$ and if $y_i < -1 + q$ then

$K(y_i) \to \infty$. The threshold $\hat{K}$ that the voter uses to reelect the incumbent is

$$\hat{K}(y_i) = \begin{cases} \infty & y_i < -1+q \\ \frac{\hat{a}(\underline{\theta})+\hat{a}(\overline{\theta})}{2} + \frac{\log\left(\frac{(1-q)(q-y_i)}{q(1-q+y_i)}\right)}{\beta(\hat{a}(\underline{\theta})-\hat{a}(\overline{\theta}))} & -1+q < y_i < q \\ -\infty & y_i > q. \end{cases}$$

Clearly, this means that if $y_i > q$ the leader is retained with probability 1 and if $y_i < -1+q$ the leader is retained with probability zero. This means that the leader's exertion of effort into mitigation policy can only affect the outcome of the election if valence is moderate, or when $-1+q < y_i < q$. Therefore, the probability of reelection can be decomposed into two terms. If $y_i > q$, the leader survives with probability 1, which occurs with $P(y_i > q) = \frac{\gamma-q}{2\gamma}$. Second, if $-1+q < y_i < q$, the leader survives with probability $\Phi(\sqrt{\beta}(a_i - \hat{K}(y_i)))$. Therefore, the total probability of survival in office is

$$\frac{1}{2\gamma}\int_{-1+q}^{q} \Phi(\sqrt{\beta}(a_i - \hat{K}(y_i)))\ dy + \frac{\gamma-q}{2\gamma}.$$

Leader $i$ maximizes the following expected utility:

$$EU_i(a_i, a_{-i}; \theta_i) = a_i + \sum_{\theta_{-i}} a_{-i}(\theta_{-i})P(\theta_{-i}) - \frac{\theta_i}{2}a_i^2 + \Big[\int_{-1+q}^{q} \Phi(\sqrt{\beta}(a_i - \hat{K}(y_i)))\ dy + \gamma - q\Big]\frac{\Psi}{2\gamma}.$$

For type $\theta_i$, the first-order condition is

$$1 - a_i\theta_i + \frac{\sqrt{\beta}\Psi}{2\gamma}\int_{-1+q}^{q} \phi\Big(\sqrt{\beta}(a_i - \frac{\hat{a}(\underline{\theta})+\hat{a}(\overline{\theta})}{2} - \frac{\log\left(\frac{(1-q)(q-y_i)}{q(1-q+y_i)}\right)}{\beta(\hat{a}(\underline{\theta})-\hat{a}(\overline{\theta}))})\Big)\ dy = 0.$$

Equilibrium requires that voters' conjectures are correct, so this simplifies to

$$1 - a_i\theta_i + \frac{\sqrt{\beta}\Psi}{2\gamma}\int_{-1+q}^{q} \phi\Big(\sqrt{\beta}(\frac{a^*(\underline{\theta})+a^*(\overline{\theta})}{2} - \frac{\log\left(\frac{(1-q)(q-y_i)}{q(1-q+y_i)}\right)}{\beta(a^*(\underline{\theta})-a^*(\overline{\theta}))})\Big)\ dy = 0.$$

Because leaders/countries are symmetric, there are 2 equations in 2 unknowns. Solving these equations yield optimal effort levels $(a^*(\underline{\theta}), a^*(\overline{\theta}))$. To confirm that the equilibrium policy choices are a maximum, I take the second-order condition. Define $\eta(a_i, y_i) = \sqrt{\beta}(a_i - \frac{\hat{a}(\underline{\theta})+\hat{a}(\overline{\theta})}{2} - \frac{\log\left(\frac{(1-q)(q-y_i)}{q(1-q+y_i)}\right)}{\beta(\hat{a}(\underline{\theta})-\hat{a}(\overline{\theta}))})$. Using the fact that

$\frac{d}{da}\phi(\eta) = -\eta\phi(\eta)\frac{\partial\eta}{\partial a}$, the second-order condition is

$$-\theta_i - \frac{\beta\Psi}{2\gamma}\int_{-1+q}^{q} \eta(a_i, y_i)\phi(\eta(a_i, y_i))\ dy.$$

Note that $\eta(a^*(\underline{\theta}), y_i) = \frac{a^*(\underline{\theta})-a^*(\overline{\theta})}{2} - \frac{\log\left(\frac{(1-q)(q-y_i)}{q(1-q+y_i)}\right)}{\beta(\hat{a}(\underline{\theta})-\hat{a}(\overline{\theta}))}) > 0$. Therefore the function inside the integral in the second-order condition for type $\underline{\theta}$ is always positive, meaning the second-order condition $-\underline{\theta} - \frac{\beta\Psi}{2\gamma}\int_{-1+q}^{q} \eta(a^*(\underline{\theta}), y_i)\phi(\eta(a^*(\underline{\theta}), y_i))\ dy < 0$ for type $\underline{\theta}$.

Now consider the second-order condition for type $\overline{\theta}$. Note that $\eta(a^*(\overline{\theta}), y_i) = \frac{a^*(\overline{\theta})-a^*(\underline{\theta})}{2} - \frac{\log\left(\frac{(1-q)(q-y_i)}{q(1-q+y_i)}\right)}{\beta(\hat{a}(\underline{\theta})-\hat{a}(\overline{\theta}))})$ need not be positive. A sufficient condition to show that the equilibrium effort $a^*(\overline{\theta})$ is a maximum is to find a lower bound on the integral. Differentiating $\eta(a^*(\overline{\theta}), y_i)\phi(\eta(a^*(\overline{\theta}), y_i))$ with respect to $y_i$ yields the critical points $y_i = \frac{q-1}{\frac{1}{qe^{\frac{1}{2}b(a^*(\overline{\theta})-a^*(\underline{\theta}))^2+\sqrt{b}(a^*(\overline{\theta})-a^*(\underline{\theta}))}-q}+1}$ and $y_i = \frac{1}{1-\frac{qe^{\frac{1}{2}b(a^*(\overline{\theta})-a^*(\underline{\theta}))^2+\sqrt{b}(a^*(\underline{\theta})-a^*(\overline{\theta}))}}{q-1}} + q - 1$. Evaluating $\eta(a^*(\overline{\theta}), y_i)\phi(\eta(a^*(\overline{\theta}), y_i))$ at the critical points yields values $-\frac{1}{\sqrt{2\pi e}}$ and $\frac{1}{\sqrt{2\pi e}}$. Further, since the integral is over an interval of length 1 with uniform density, the integral has a lower bound of $-\frac{1}{\sqrt{2\pi e}}$. Substituting this into the second-order condition yields the condition

$$-\overline{\theta} + \frac{\beta\Psi}{2\gamma}\frac{1}{\sqrt{2\pi e}} \leq 0,$$

yielding the condition $\beta \leq \frac{2\gamma\overline{\theta}\sqrt{2\pi e}}{\Psi}$.

Since the second-order condition is negative at the equilibrium effort choice, it is a maximum. Further, this is the only maximum by concavity of the utility function. Therefore, such an optimal policy must be unique. Indeed, this is the unique equilibrium because pooling equilibria cannot exist. Pooling can be ruled out by noticing that, in any pooling equilibrium, the probability of reelection is not a function of the choice variable (i.e, it is a constant). The solution to the problem in that case is the leader's ideal point, $\tilde{a}(\theta_i) = \frac{1}{\theta_i}$ in which $\tilde{a}(\underline{\theta}) \neq \tilde{a}(\overline{\theta})$, contradicting pooling. ■

**Claim 2** *Leader i's equilibrium policy choices are decreasing in $\theta$, increasing in $\Psi$, and decreasing in $\gamma$.*

**Proof of Claim 2:** Observe that the leader's utility function exhibits decreasing differences with respect to $a$ and $\theta$ for all values of $a$ and $\theta$ (analogous for $a$ and $\gamma$). Additionally, the utility function exhibits increasing differences with respect to $a$ and $\Psi$ for all values of $a$ and $\Psi$. Therefore, by the tools of monotone comparative statics (Ashworth and Bueno de Mesquita 2006), I conclude that $\frac{\partial a_i^*}{\partial\theta_i} \leq 0$, $\frac{\partial a_i^*}{\partial\gamma} \leq 0$, and $\frac{\partial a_i^*}{\partial\Psi} \geq 0$ for any $a_i^*$ that maximizes leader $i$'s expected utility.

We have for any $\theta$ and any $\Psi$ at the equilibrium choice of $a_i$,

$$\frac{\partial^2 EU_i}{\partial a_i \partial \theta_i} = -a_i < 0.$$
$$\frac{\partial^2 EU_i}{\partial a_i \partial \Psi} = \frac{\sqrt{\beta}}{2\gamma} \int_{-1+q}^{q} \phi\Big(\sqrt{\beta}(a_i - \hat{K}(y_i))\Big)\, dy > 0.$$
$$\frac{\partial^2 EU_i}{\partial a_i \partial \gamma} = -\frac{\sqrt{\beta}\Psi}{2\gamma^2} \int_{-1+q}^{q} \phi\Big(\sqrt{\beta}(a_i - \hat{K}(y_i))\Big)\, dy < 0.$$

■

**Proof of Proposition 1:** Existence and uniqueness of the equilibrium is established in Claim 1. The ideal effort choice can be found by differentiating $A - \frac{\theta_i}{2}a_i^2$ and solving for $a_i$, yielding the maximum $\tilde{a}_i = \frac{1}{\theta_i}$. That $a^*(\underline{\theta}) > a^*(\overline{\theta})$ follows from Claim 2 because policy choices are increasing in $\theta$. That $\frac{\partial a_i^*}{\partial \Psi} \geq 0$ is immediate from Claim 2. Since the probability of surviving is increasing in $K_i$ and $K_i$ is increasing in effort $a_i$, congruent leaders are likely to survive in office than incongruent leaders because $a^*(\underline{\theta}) > a^*(\overline{\theta})$. ■

**Lemma 1** *Suppose $\Psi = 0$. Any incentive compatible mechanism is "compressed:" $x(\underline{\theta}) = x(\overline{\theta})$. The recommendation takes the form*

$$x^*(\hat{\theta}_i) = \frac{n}{q\underline{\theta} + (1-q)\overline{\theta}}.$$

**Proof of Lemma 1:** Since types are private information, the IO's objective function is

$$V = \max_{\{x(\underline{\theta}),\ x(\overline{\theta})\}} \sum_i q\Big[x(\underline{\theta}) + \sum_{\theta_{-i}} X_{-i}(\theta_{-i})P(\theta_{-i}) - \frac{\underline{\theta}}{2}x(\underline{\theta})^2\Big] + (1-q)\Big[x(\overline{\theta}) + \sum_{\theta_{-i}} X_{-i}(\theta_{-i})P(\theta_{-i}) - \frac{\overline{\theta}}{2}x(\overline{\theta})^2\Big].$$

The IO wishes to maximize $V$ subject to the incentive constraints

$$x(\theta_i) + \sum_{\theta_{-i}} X_{-i}(\theta_{-i})P(\theta_{-i}) - \frac{\theta_i}{2}x(\theta_i)^2 \geq x(\theta_i') + \sum_{\theta_{-i}} X_{-i}(\theta_{-i})P(\theta_{-i}) - \frac{\theta_i}{2}x(\theta_i')^2.$$

The monotonicity of the incentive constraints requires that $x(\underline{\theta}) \geq x(\overline{\theta})$. Notice that the incongruent type would never mimic the congruent type, as doing so would lead her to receive a more stringent recommendation than she would prefer. The congruent type, however, could choose to mimic the incongruent type, receiving a less ambitious recommendation. Therefore, the incentive constraint of the congruent type must bind.

Rewriting it slightly implies that

$$x(\underline{\theta}) - \frac{\underline{\theta}}{2}x(\underline{\theta})^2 = x(\overline{\theta}) - \frac{\underline{\theta}}{2}x(\overline{\theta})^2 + (\frac{\overline{\theta}}{2} - \frac{\underline{\theta}}{2})x(\overline{\theta})^2.$$

Substituting this into the IO's objective function yields

$$\begin{aligned} V &= \sum_i \sum_{\theta_{-i}} X_{-i}(\theta_{-i})P(\theta_{-i}) + q\Big[x(\overline{\theta}) - \frac{\underline{\theta}}{2}x(\overline{\theta})^2 + (\frac{\overline{\theta}}{2} - \frac{\underline{\theta}}{2})x(\overline{\theta})^2\Big] + (1-q)\Big[x(\overline{\theta}) - \frac{\overline{\theta}}{2}x(\overline{\theta})^2\Big] \\ &= \sum_i \sum_{\theta_{-i}} X_{-i}(\theta_{-i})P(\theta_{-i}) + x(\overline{\theta}) - q\frac{\underline{\theta}}{2}x(\overline{\theta})^2 - (1-q)\frac{\overline{\theta}}{2}x(\overline{\theta})^2. \end{aligned}$$

Since countries are symmetric, it must be the case that the incentive constraint for all competent types bind simultaneously, so $\sum_{\theta_{-i}} X_{-i}(\theta_{-i})P(\theta_{-i}) = (n-1)x(\overline{\theta})$. Therefore,

$$V = n\Big[nx(\overline{\theta}) - q\frac{\underline{\theta}}{2}x(\overline{\theta})^2 - (1-q)\frac{\overline{\theta}}{2}x(\overline{\theta})^2\Big].$$

The solution to this problem is compressed. This means that the IO assigns the same policy regardless of reported type, $x^*(\underline{\theta}) = x^*(\overline{\theta})$. Such a policy is incentive compatible because it yields the same utility regardless of whether leader $i$ reports $\hat{\theta}_i = \underline{\theta}$ or $\hat{\theta}_i = \overline{\theta}$. To see this, notice that if not, $x^*(\underline{\theta}) \neq x^*(\overline{\theta})$, monotonicity requires $x(\underline{\theta}) > x(\overline{\theta})$ for $\underline{\theta} < \overline{\theta}$. Finally, because of the concavity of the leader's utility function, we have that for $\underline{\theta} < \overline{\theta}$, the congruent type's interim expected utility is greater if it mimics the incongruent type, $x(\overline{\theta}) - \frac{\underline{\theta}}{2}x(\overline{\theta})^2 > x(\underline{\theta}) - \frac{\underline{\theta}}{2}x(\underline{\theta})^2$, which contradicts incentive compatibility. Thus any solution is compressed.

Since $x^*(\underline{\theta}) = x^*(\overline{\theta}) = x^*$, the solution to this maximization problem is $x = \frac{n}{q\underline{\theta} + (1-q)\overline{\theta}}$, as in the lemma. ■

**Lemma 2** *No incentive compatible mechanism exists if* $\Psi > \frac{\gamma(\overline{\theta}\omega - 1)^2}{\overline{\theta}}$.

**Proof of Lemma 2:** The incentive compatibility constraint for the incongruent type can be written as

$$x(\overline{\theta}) - \frac{\overline{\theta}}{2}x(\overline{\theta})^2 - x(\underline{\theta}) + \frac{\overline{\theta}}{2}x(\underline{\theta})^2 - \frac{1}{2\gamma}\Psi \geq 0.$$

Clearly, this establishes an upper bound on $\Psi$ as in the text, $\Psi \leq \gamma(x(\underline{\theta}) - x(\overline{\theta}))(\overline{\theta}(x(\underline{\theta}) + x(\overline{\theta})) - 2)$. We wish to maximize this bound so that $\Psi$ can fall within as large a range as possible. Define $IC(x(\underline{\theta}), x(\overline{\theta})) = x(\overline{\theta}) - \frac{\overline{\theta}}{2}x(\overline{\theta})^2 - x(\underline{\theta}) + \frac{\overline{\theta}}{2}x(\underline{\theta})^2 - \frac{1}{2\gamma}\Psi$. We wish to maximize this function on the domain $x(\theta_i) \in [0, \omega]$

subject to the requirement that $x(\underline{\theta}) > x(\overline{\theta})$. The maximum thus occurs at $x(\underline{\theta}) = \omega$ and $x(\overline{\theta}) = \frac{1}{\overline{\theta}}$ such that $IC(\omega, \frac{1}{\overline{\theta}}) = \frac{(\overline{\theta}\omega-1)^2}{2\overline{\theta}} - \frac{1}{2\gamma}\Psi$. Thus, for incentive compatibility to hold, we require that $\Psi \leq \frac{\gamma(\overline{\theta}\omega-1)^2}{\overline{\theta}}$. If $\Psi > \frac{\gamma(\overline{\theta}\omega-1)^2}{\overline{\theta}}$, then the incongruent type's incentive constraint is never satisfied, regardless of how the IO designs $x(\theta_i)$. ■

**Lemma 3** *Recommendations $x^*(\theta_i)$ satisfy leaders' obedience constraints if and only if $x^*(\theta_i) = \frac{1}{\theta_i} = \tilde{a}(\theta_i)$.*

**Proof of Lemma 3:** The obedience constraint of leader $i$ with type $\theta_i$ is

$$x(\theta_i) + \sum_{\theta_{-i}} X_{-i}(\theta_{-i})P(\theta_{-i}) - \frac{\theta_i}{2}x(\theta_i)^2 + \frac{\gamma - q + \mu}{2\gamma}\Psi \;\geq\; \max_d \; d + \sum_{\theta_{-i}} X_{-i}(\theta_{-i})P(\theta_{-i}) - \frac{\theta_i}{2}d^2 + \frac{\gamma - q + \mu}{2\gamma}\Psi.$$

The solution to the right-hand side is that the optimal deviation is $d = \frac{1}{\theta_i}$. Substituting in $d = \frac{1}{\theta_i}$ yields

$$x(\theta_i) - \frac{\theta_i}{2}x(\theta_i)^2 - \frac{1}{2\theta_i} \geq 0,$$

which requires that $x(\theta_i) = \frac{1}{\theta_i}$ and the constraint is met with equality. ■

**Proposition 2** *In an equilibrium implemented by the international cooperation game:*

- *leaders invest effort at their ideal points, $x^*(\theta_i) = \tilde{a}(\theta_i) = \frac{1}{\theta_i}$, only if $\Psi \in [0, \frac{\gamma(\overline{\theta}-\underline{\theta})^2}{\overline{\theta}\underline{\theta}^2}]$;*
- *congruent leaders invest greater effort into mitigation than incongruent leaders, $x^*(\underline{\theta}) > x^*(\overline{\theta})$;*
- *the value of office-holding has no effect on mitigation efforts, $\frac{\partial}{\partial\Psi}x^*(\theta_i) = 0$;*
- *congruent leaders are more likely to survive in office than incongruent leaders;*
- *expected aggregate efforts are less than those in the domestic politics game, $E[X^*] \leq E[A^*]$.*

**Proof of Proposition 2:** That the implementable policy is the ideal effort follows from Lemma 3, since it is the only policy that would satisfy obedience constraints. Substituting $x(\theta_i) = \frac{1}{\theta_i}$ into the congruent leader's incentive constraints yields $\Psi \geq -\frac{\gamma(\overline{\theta}-\underline{\theta})^2}{\overline{\theta}^2\underline{\theta}}$, which always holds, implying that implementing ideal points is always incentive compatible for the congruent type. For the incongruent type, substituting yields $\Psi \leq \frac{\gamma(\overline{\theta}-\underline{\theta})^2}{\overline{\theta}\underline{\theta}^2}$, establishing the upper bound on $\Psi$ for ideal points to be incentive compatible. It is immediate that $\frac{1}{\underline{\theta}} > \frac{1}{\overline{\theta}}$ and $\frac{\partial}{\partial \Psi}\frac{1}{\theta_i} = 0$. Finally, aggregate expected emissions in the international cooperation game are $E[X^*] = n(\frac{q}{\underline{\theta}} + \frac{1-q}{\overline{\theta}})$, and aggregate expected emissions in the domestic politics game are $n(qa^*(\underline{\theta}) + (1-q)a^*(\overline{\theta}))$. By Claim 2, the equilibrium policies of the domestic politics game are greater than the ideal policies. ■

**Proposition 3** *If leaders are expected to be incongruent $q < \frac{1}{2}$, there exists a threshold $\overline{\gamma}$ such that leaders join the agreement whenever $\gamma > \overline{\gamma}$.*

**Proof of Proposition 3:** Leader $i$'s expected utility for joining the agreement is

$$EU_i(\text{join}) = q\Big[\frac{1}{\underline{\theta}} + E[X^*_{-i}] - \frac{1}{2\underline{\theta}} + \frac{\gamma - q + 1}{2\gamma}\Psi\Big] + (1-q)\Big[\frac{1}{\overline{\theta}} + E[X^*_{-i}] - \frac{1}{2\overline{\theta}} + \frac{\gamma - q}{2\gamma}\Psi\Big].$$

Define $p(\theta_i) = \int_{-1+q}^{q} \Phi(\sqrt{\beta}(a^*(\theta_i) - \hat{K}^*(y_i)))\ dy$. Note that, by Claim 2, $\frac{\partial p(\theta_i)}{\partial \gamma} \leq 0$ because $\frac{\partial a^*(\theta_i)}{\partial \gamma} \leq 0$.

If leader $i$ does not join the agreement, her expected utility is

$$EU_i(\neg\text{join}) = q\Big[a^*(\underline{\theta}) + E[A^*_{-i}] - \frac{\underline{\theta}}{2}a^*(\underline{\theta})^2 + \frac{\gamma - q + p(\underline{\theta})}{2\gamma}\Psi\Big] + (1-q)\Big[a^*(\overline{\theta}) + E[A^*_{-i}] - \frac{\overline{\theta}}{2}a^*(\overline{\theta})^2 + \frac{\gamma - q + p(\overline{\theta})}{2\gamma}\Psi\Big].$$

By symmetry, if leader $i$ joins, then all leaders join, so $E[X_{-i}] = (n-1)(\frac{q}{\underline{\theta}} + \frac{1-q}{\overline{\theta}})$. Similarly, if leader $i$ does not join, then everyone stays out of the agreement, so $E[A_{-i}] = (n-1)(qa^*(\underline{\theta}) + (1-q)a^*(\overline{\theta}))$. Substituting these quantities into the expected utilities yields the expected utility that any leader would or would not join the agreement (since all leaders are symmetric conditional on type). Therefore the expected utilities from joining and not joining are respectively

$$EU(\text{join}) = \frac{(2n-1)(q(\overline{\theta} - \underline{\theta}) + \underline{\theta})}{2\overline{\theta}\underline{\theta}} + \frac{\Psi}{2}.$$

$$EU(\neg\text{join}) = \frac{1}{2}\Big(qa^*(\underline{\theta})(2n - \underline{\theta}a^*(\underline{\theta})) + (1-q)a^*(\overline{\theta})(2n - \overline{\theta}a^*(\overline{\theta})) + \Psi + \frac{\Psi((1-q)p(\overline{\theta}) - q(1-p(\underline{\theta})))}{\gamma}\Big).$$

The expected utility of joining is constant in $\gamma$. Differentiating the expected utility of not joining with respect to $\gamma$ yields

$$\frac{\partial EU(\neg\text{join})}{\partial \gamma} = (1-q)a^*(\overline{\theta})(n - \overline{\theta}a^*(\overline{\theta}))\frac{\partial a^*(\overline{\theta})}{\partial \gamma} + qa^*(\underline{\theta})(n - \underline{\theta}a^*(\underline{\theta}))\frac{\partial a^*(\underline{\theta})}{\partial \gamma} + \frac{\Psi}{2\gamma^2}\Big(\gamma q\frac{\partial p(\underline{\theta})}{\partial \gamma} + \gamma(1-q)\frac{\partial p(\overline{\theta})}{\partial \gamma} + q - (1-q)p(\overline{\theta}) - qp(\underline{\theta})\Big).$$

The first two terms are negative because $\frac{\partial a^*(\theta_i)}{\partial \gamma} \leq 0$ by Claim 2. The final term will always be negative if $q < \frac{p(\overline{\theta})}{1 - p(\underline{\theta}) + p(\overline{\theta})}$, which is always at least as large as $\frac{1}{2}$. Thus a sufficient condition for the final term to be negative is if $q < \frac{1}{2}$. Therefore by the intermediate value theorem there exists a threshold $\overline{\gamma}$ such that leaders prefer to join whenever $\gamma > \overline{\gamma}$. $\blacksquare$