Designing the Optimal International Climate Agreement with Variability in Commitments*

Jordan H. McAllister[†] and Keith Schnakenberg[‡]

May 28, 2021

Abstract

We analyze the design of an international climate agreement. In particular, we consider two goals of such an agreement: overcoming free-rider problems and adjusting for differences in mitigation costs between countries. Previous work suggests that it is difficult to achieve both of these goals at once under asymmetric information because countries free ride by exaggerating their abatement costs. We argue that independent information collection (investigations) by an international organization can alleviate this problem. In fact, though the best implementable climate agreement without investigations fails to adjust for individual differences even with significant enforcement power, a mechanism with investigations allows for adjustment and can enable implementation of the socially optimal agreement. Furthermore, when the organization has significant enforcement power, the optimal agreement is achievable even with minimal investigative resources (and vice versa). The results suggest that discussions about institutions for climate cooperation should focus on information collection as well as enforcement.

*The authors thank Amanda Kennard and David Victor as well as participants in the Washington University formal theory seminar for valuable comments.

[†]Department of Political Science, Washington University in St. Louis, jordan.mcallister@wustl.edu

[‡]Department of Political Science, Washington University in St. Louis, keschnak@wustl.edu

Introduction

The international community has failed to significantly slow the pace of climate change even though its causes are known and its effects are potentially devastating.¹ This failure is partially explained by incentives: individual governments are tempted to let other countries bear the costs of climate change abatement and reap the benefits without taking on these costs themselves.² As a result, scholars and practitioners have emphasized the importance of well-designed international institutions to help shape the incentives of national actors.³ In so doing, an ideal institution would need to not only manage free riding, but also reflect variability between countries by asking countries that can reduce fossil fuel emissions at a lower cost to do more and requiring others to do less.⁴

The importance of accounting for individual differences has been recognized in major efforts to create a global climate agreement. First, the precursor to modern climate agreements—the 1987 Montreal Protocol on the ozone layer—set the standard by assigning non-uniform commitments in which developing countries were given less strict emissions limits.⁵ This was echoed in later climate agreements such as the 1992 United National Framework Convention on Climate Change (UNFCCC), which recognized the “common but differentiated responsibilities” of all countries.⁶ In due course, the 1997 Kyoto Protocol set emissions limits only for developed countries and even set country-specific limits wherein economies in transition were allowed to use a different base year by which to compare their reductions.⁷ Most recently, the 2016 Paris Agreement simply let each country determine its own quota.⁸ Agreements with non-uniform obligations are more cost-effective because member states have varying costs, as well as broader because variability ensures that each country’s quota is appropriate given the country’s relative capability.⁹

Reflecting variability between countries is a desirable component of a climate agreement. First, a one-size-fits-all policy would be suboptimal because some countries are able to reduce fossil fuel emissions at a much lower cost than others. In an agreement with uniform requirements, some

¹Stern (2007)

²Barrett and Nitze (2007); Bernauer (2013); Clapp and Dauvergne (2011); Hovi, Sprinz, and Underdal (2009); Tavoni et al. (2011)

³Martin (1992); Michaelowa and Michaelowa (2017); Ostrom (1990)

⁴Keohane and Victor (2011); McGinty (2007)

⁵Barrett (2003)

⁶Stern (2007)

⁷Barrett (2003); Von Stein (2008)

⁸Harstad (2018); Hovi et al. (2019)

⁹Barrett (2003)

countries bear disproportionate burdens while others that could do more are not required to do so.¹⁰ Second, one-size-fits-all agreements suffer from a “broad versus deep” trade-off which limits their success. If international agreements require the same commitments of all members, then they can either have broad membership or require deep commitments, but not both.¹¹ By contrast, a variable agreement encourages both broader participation and deeper commitments for countries able to make them by setting each country’s quota separately according to each nation’s capabilities. However, because countries’ leaders are likely more informed about the individual circumstances within their own countries than are leaders of other countries or of international organizations, an agreement must rely to some extent on self-reports—relatedly, it could be the case that an international agency such as the United Nations Environment Programme (UNEP) gathers information transmitted by countries prior to drafting an agreement (as in the Montreal Protocol), or that an agreement relies on each country to determine the level of commitment it is able to make (as in the Paris Agreement)—to implement a variable agreement.¹²

Unfortunately, as Harrison and Lagunoff (2017) have argued, overcoming free riding while accounting for individual differences is difficult in the face of this asymmetric information. To see why, consider an example of an agreement involving only two countries and a mediator helping countries decide what commitments they must make under the agreement. To set enforcement issues aside for the moment, assume the two countries have significant enough repeated interactions that they can make credible commitments to one another, knowing that they can mutually punish defection from their agreement.¹³ The ideal agreement has a country commit to only a moderate reduction in emissions when it faces a hardship resulting from a change in economic or political conditions, and to a more significant reduction when it faces no such hardship.¹⁴ Consistent with the free-riding problem, however, a country facing no hardship may still prefer to commit to only moderate reductions and let the other country carry more of the costs of climate cooperation. This creates a problem for the mediator trying to decide what commitments each country should make

¹⁰See Bernauer (2013). Though we are referring solely to efficiency when we discuss disproportionate burdens as suboptimal, one-size-fits-all agreements may be especially unacceptable to climate negotiators with preferences for equity, as in Dannenberg, Sturm, and Vogt (2010) and Kennard (2020).

¹¹Barrett (1994); Barrett (2003)

¹²The Montreal Protocol, UNFCCC, Kyoto Protocol, and Paris Agreement all rely on member states’ self-reports for information collection (Barrett 2003; Hovi et al. 2019; Slechten 2020).

¹³Hovi, Ward, and Grundig (2015); Milgrom, North, and Weingast (1990); Rosendorff and Milner (2001)

¹⁴Koremenos (2005); Pelc (2009)

under information asymmetry.¹⁵ Each country is tempted to exaggerate hardships so that it is allowed to make lower commitments under the agreement.¹⁶ Information asymmetry is significant enough that in some circumstances, the best agreement that can be implemented is one that contains no variation in commitments.¹⁷

In this paper we analyze how to design an optimal climate agreement when countries' costs of climate action are private information. The problem that the international organization (IO) must solve is to maximize global welfare subject to the constraint that countries are willing to participate and also to submit accurate self-reports. To focus on the informational aspect of the problem, we assume that the international organization has some (perhaps limited) ability to enforce commitments that member states make.¹⁸ For instance, enforcement leverage may come from some positive linkage such as controlling access to a club good or other benefit that accrues to countries for being in good standing with the organization,¹⁹ or even from some negative linkage such as punishments used to coerce countries into compliance.²⁰ Nonetheless, without any investigative resources, this leverage for inducing compliance does not help the organization design an optimal agreement.

Even as the IO has very high leverage for enforcement, it cannot prevent countries from exaggerating compliance costs. However, when the IO has the power to investigate the compliance costs of member states, it may be able to overcome misreporting while also allowing variability in commitments. Notably, the availability of investigative resources means that the IO's leverage for encouraging compliance can also help it incentivize truth telling to make it so that the agreement is able to assign realistic commitments for each member state. This could be a benefit of major climate agreements with a verification procedure to check countries' reports, including the Montreal Protocol (for the ozone), UNFCCC, Kyoto Protocol, and Paris Agreement.²¹ To increase their investigative

¹⁵See also Slechten (2020).

¹⁶See Herzing (2005), Karp and Zhao (2010), and Slechten (2020) for discussions of how the unobserved nature of economic shocks, abatement costs, or domestic opposition produces an incentive problem in which countries overestimate the size of any such barriers to lowering their commitments. This also mirrors the standard problem of mechanism design for public goods provision where agents have private information about their willingness to pay for the public good. See, for example Csapó and Müller (2013).

¹⁷Harrison and Lagunoff (2017)

¹⁸This simplifies the exposition of results but is easily microfounded. In the Appendix, we discuss endogenous enforcement power through dynamic agreements or issue bundling. In both cases, punishments or rewards are self-enforcing and do not require direct policing power by the IO.

¹⁹Keohane and Victor (2016); Nordhaus (2015); Stern (2007)

²⁰Barrett (2003); Martin (1992); Mitchell and Keilbach (2001)

²¹See, for instance, Barrett (2003), Hovi et al. (2019), and Slechten (2020) for descriptions of these procedures.

capacity, IOs can also oversee investigations conducted by external bodies. Indeed, although most environmental organizations lack significant monitoring capabilities, they have historically been able to rely on NGOs or other actors to conduct monitoring for them. For example, the 1979 Long-Range Transboundary Air Pollution Convention did not have the ability to monitor but was able to depend on environmental NGOs to do so.²² Likewise, because free-riding behavior is so costly to all countries in an agreement, then they will have the incentive to request or acquire extensive information on each other’s true types.²³ So even if the organization itself lacks the ability to collect all of the information it needs, it can still expect entities such as NGOs and member states to conduct their own investigations.

We contribute to the literature in two main ways. First, we contribute to the literature on mechanism design and climate change cooperation. In general, we follow the literature in viewing the main function of climate agreements as attempting to overcome problems of free riding.²⁴ The closest papers are Harrison and Lagunoff (2017) and Slechten (2020). Harrison and Lagunoff (2017) take a mechanism design approach to constructing an international agreement to limit carbon consumption. In their model, the best implementable agreement subjects every country to the same requirements even though the best possible policy would account for individual differences in costs and abilities. The problem is that if countries have private information about these differences, then they may try to free ride by exaggerating their costs from reducing carbon consumption. We complement that paper by considering independent information gathering as a potential solution to this issue of misreporting. Slechten (2020) also considers the possibility of verifying self-reports in a mechanism design setting for environmental agreements. In her model, states can voluntarily allow the IO to “certify” their private information, making it completely verifiable. By contrast, we consider probabilistic verification and allow the IO to freely allocate investigative resources determining the likelihood of verification. We also differ from Slechten (2020) by considering a multilateral agreement with no transfers, reflecting a weak IO that cannot freely move resources between countries.

Second, we contribute to the international relations literature on flexibility in the design of

Though these procedures are often used to check reports of compliance, we suggest that similar institutions could in theory be employed for wider information collection, as depicted in the model by Slechten (2020).

²²Dai (2007)

²³Martin (1992)

²⁴Hovi, Ward, and Grundig (2015); Nordhaus (2015); Ostrom (1990)

international institutions more broadly. Much of the work on flexibility comes from the literature on trade agreements such as the World Trade Organization (WTO).²⁵ In these papers, two factors driving the need for flexibility are changes over time that require renegotiation²⁶ and individual differences between countries leading to a need for escape clauses²⁷ or compensatory payments.²⁸ We conceptualize flexibility in agreements differently than in these papers, and use the term *variability* instead to avoid conflating these ideas. Though these papers refer either to renegotiation or to forgiveness from punishment, we simply mean allowing the terms of the agreement to be different for each country. However, some of the basic insights about flexibility are similar, and we believe similar informational problems to those raised in this paper could apply to escape clauses, compensation, or renegotiation.²⁹ The flexible mechanism closest to the kind of variability in this paper is the escape clause, which enables countries to temporarily lower their commitments when their costs of abatement (or type) become higher due to an economic shock, military conflict, etc.³⁰

In particular, our paper speaks to the climate agreements literature’s commonly noted broad versus deep trade-off between the strictness of actions required under an agreement and the ability to gain more signatories.³¹ The basic intuition behind this trade-off is that an agreement that requires more significant action on the part of its members will therefore be more costly for those countries, making it so that all countries will be less inclined to sign on to the agreement initially.³² In other words, an agreement with uniform commitments can either be small in membership but assign a high quota (in which case only the most-committed countries would be members), or large in membership but assign a low quota (in which case only the least-committed countries would be curbing their emissions). This problem is particularly difficult when it comes to climate change because non-members of climate agreements are not only refraining from action, but they are also making the problem worse because maintaining business as usual actually results in a further

²⁵See Goldstein and Martin (2000); Herzing (2005); Pelc (2009); Rosendorff (2005); Rosendorff and Milner (2001); Sykes (1991). See also Gilligan and Johns (2012) for a useful review of some of this literature.

²⁶Barrett (2003); Harstad (2018); Koremenos (2001); Kucik and Reinhardt (2008)

²⁷Helfer (2012); Koremenos (2005); Koremenos, Lipson, and Snidal (2001); Rosendorff and Milner (2001); Sykes (1991)

²⁸Herzing (2005); Kucik and Reinhardt (2008)

²⁹Boockmann and Thurner (2006)

³⁰See Pelc (2009). Regarding environmental agreements in particular, Karp and Zhao (2010) show how allowing a country that exceeds its emissions quota to use an escape clause enables an environmental agreement to simultaneously achieve efficient abatement and high participation. In fact, around one-fourth of environmental agreements contain an escape clause (Marcoux 2009).

³¹Hovi, Ward, and Grundig (2015); Keohane and Victor (2016); Von Stein (2008)

³²Barrett (1994); Barrett (2003)

deterioration of the climate. Therefore, it is not evident whether it is better to try to craft a deeply committed agreement of only the most willing countries and risk letting less-willing countries pollute as normal, or to form a weak agreement that includes all countries but which does not require much of any of its members.³³ Variability in commitments may increase the number of willing signatories by ensuring that countries that face higher compliance costs receive some relief. An agreement that allows variation in the requirements made of less well-off countries or in less secure times will be able to both require more of countries and get more countries to sign on, because these countries know they will not be asked to do more than they are able at any given time.³⁴ Our conception of variation in this paper is the extent to which the agreement's requirements account for individual differences. An optimally varying agreement softens the trade-off between breadth and depth but requires independent investigations of member states' claims to overcome the incentive of countries to exaggerate costs of compliance. We now turn to the model.

The game

In the benchmark model, the players comprise N countries, where $N = \{1, \dots, n\}$. We assume that each of these countries $i \in N$ is able to choose its own level of carbon emissions $c_i \geq 0$, and that all countries decide on this simultaneously.³⁵ However, we also assume that each country has a type θ_i which takes a value in $[\underline{\theta}, \bar{\theta}] \subset (0, 1)$. Higher types value carbon consumption more than other countries or would incur relatively high costs from lowering emissions, while lower types value abatement more than other countries or would incur relatively low costs from lowering emissions. For example, a high type of country could be one that will adapt easily to climate change or whose economic strategy entails continued investment in energy-intensive manufacturing sectors. A low type of country may be one invested in clean technologies or less energy-intensive industries. We may also think of types as related to internal political pressures within a country, given that the policy preferences of national leaders are influenced by the relative power of environmental interest

³³Our benchmark model characterizes the optimal *fully inclusive* mechanism, meaning that all countries are willing to sign on, but we do include in the Appendix an example showing that the optimal compressed mechanism need not be fully inclusive and discuss the structure of non-inclusive mechanisms with no investigations. Once we introduce investigations, the optimal mechanism is always fully inclusive.

³⁴Barrett (2003); Rosendorff (2005); Rosendorff and Milner (2001); Von Stein (2008)

³⁵In reality, countries may choose policies such as taxes or subsidies that indirectly regulate the aggregate amount of carbon emissions by private actors. For the sake of simplicity, we abstract away from these choices and act as if the country can simply choose the level of carbon emissions.

groups, industry lobbyists, and so on. Types are private information but are commonly known to be independent and identically distributed along a continuous distribution F with density f .

Based on these assumptions, this model then focuses on the possible strategies that could be employed by an international organization to guide countries' compliance with global abatement goals. In particular, the IO is able to first collect information about countries' types by asking countries to voluntarily disclose this information through their announcements $\hat{\theta}_i$, and then it can recommend an emissions quota for each country. We denote this quota by $\tilde{c}(\hat{\theta}_i)$ to show that it may depend on type announcements. The IO in our model is a utilitarian planner which seeks to maximize the sum of all countries' payoffs.³⁶

Each country's payoff is a function of how much carbon is emitted and its type, which can be written as follows:

$$u_i(c, \theta_i) = \theta_i \log(c_i) + (1 - \theta_i) \log(\omega - C). \quad (1)$$

In this expression $C = \sum_{i=1}^n c_i$ is the aggregate level of carbon emissions and $\omega > 0$ represents the carbon stock. If country i does not sign on to the agreement, it receives $u_i(\mathbf{c}, \theta_i)$ for its chosen level of emissions but pays a fixed penalty $K \geq 0$ for noncooperation. The first term in the utility function represents the countries' direct benefits from carbon consumption, while the second term represents the benefits from conservation. The type θ_i serves as a weight on these two components, so naturally countries value conservation less relative to consumption as θ_i increases.

This form captures the idea that high types' utility increases by more when they emit (because the utility function is written in a way such that emissions are valued relatively more and conservation relatively less when the type is high), while low types' utility increases by less when they emit (because the utility function is written so conservation increases in value and emissions decrease in value when the type is low). For a country that has not signed on to the agreement, it has this same utility function related to emissions and abatement but also pays a fixed penalty of K

³⁶Our assumption of a utilitarian IO is primarily motivated by our research question. We want to know whether it is possible to implement a socially optimal agreement, so we take as given that the IO would want to implement such an agreement if it could. As a consequence, we assume away agency loss from delegating to the IO even though this may be empirically relevant. For instance, if the IO sees its principal as primarily consisting of a small group of powerful countries (e.g., Copelovitch (2010)) then the IO's actual objective may differ significantly from the one we use here. These considerations are interesting avenues for future research, but we put them aside in this paper to preserve clarity on the question of whether socially optimal agreements are possible.

for noncooperation. This fixed cost could be seen as the loss of any club benefits,³⁷ alliances, or simply prestige as a result of not being in the agreement. For the moment, we will treat K as exogenous. However, in the Appendix we consider an extension in which K arises endogenously from issue bundling. One thing we wish to make clear is that the IO need not have enforcement power in a formal sense (as these organizations often do not). Instead, fairly large values of K can be sustained by the IO just helping countries to coordinate on self-enforcing equilibria that punish climate defectors.

Overall, we analyze an institution in which each country makes an announcement to the IO and the IO recommends an emissions level. Though this seems restrictive if it is taken literally, the *Revelation Principle* implies that the outcome of any equilibrium from any institution can be implemented by a direct mechanism such as this one in which all players have an incentive to truthfully report their types. That is, suppose an outcome (i.e., emissions levels for each type) is implemented as an equilibrium to some potentially complicated mechanism. The Revelation Principle tells us that the same outcome can be implemented by a direct mechanism in which countries are asked to reveal their types and then are assigned emissions levels. Therefore, our setup encompasses a very broad set of institutions.³⁸

The form of the utility functions and the choices of the IO in the benchmark model follow previous work by Harrison and Lagunoff (2017). The primary differences are that the model in Harrison and Lagunoff (2017) is dynamic in contrast to our static model, and that our model imposes a positive cost of noncooperation on countries that fall outside of the agreement. These two choices are connected: the dynamic Harrison and Lagunoff (2017) model generates punishments for noncompliance endogenously, but our static model requires an exogenous punishment for noncooperation. Our setup allows for simpler analysis and also lets us compare what happens when the IO can impose large costs of noncooperation versus when it cannot.³⁹

³⁷Nordhaus (2015)

³⁸The fact that any implementable mechanism *can be* implemented by a direct mechanism does not imply that we should only observe direct mechanisms in the world. There may be political or practical reasons to use procedures that do not simply involve direct self-reports followed by quota recommendations. Our point is to analyze what can be achieved. The revelation principle suggests that analyzing truthful direct mechanisms is a useful way to do that, because any equilibrium of any game with these preferences and types is outcome-equivalent to some truthful direct mechanism.

³⁹Additionally, in a dynamic model the IO may not benefit from imposing large costs of noncompliance because doing so reduces the ability of low types to credibly threaten to stop cooperating if other countries defect (Cirone and Urpelainen 2013). This suggests that a dynamic model with investigations is an interesting avenue for future research.

Complete information benchmark

First, we solve for the quotas the IO would suggest assuming that there was perfect information about countries' types and that there was no need to deal with any perverse incentives held by the countries. This complete information benchmark has a socially optimal quota which is increasing in the global carbon stock and each country's type but decreasing in the number of countries, as suggested by the following lemma:

Lemma 1. *Under complete information the socially optimal quota is $\tilde{c}(\theta_i) = \frac{\omega\theta_i}{n}$*

Lemma 1 implies that those countries that value their emissions more are able to emit more under this quota (because the optimal quota is increasing in θ_i), but the quotas for all countries are also influenced by how many countries are a part of the agreement (because the extent of this allowance is decreasing in the total number of countries and increasing in the total available carbon stock).

Our complete information benchmark captures the idea that the optimal policy is variable in the sense of allowing for individual differences. In the next section we introduce private information and reproduce the full compression result of Harrison and Lagunoff (2017) which states that the best implementable quota is completely inflexible, requiring the same actions by all countries.

Best mechanism under constraints

Our next step is to consider how an IO could attempt to implement an agreement with some limited enforcement power but no investigative powers. We thus consider the quotas the IO would suggest after imposing countries' incentive constraints. The first of these two types of constraints is the incentive compatibility constraint, which says that each country must find it in its best interest to reveal its true type given the quotas associated with each announced type:

$$\int_{\Theta_{-i}} u_i(\tilde{c}(\theta_i), \tilde{\mathbf{c}}(\theta_{-i}), \theta_i) dF_{-i}(\theta_{-i}) \geq \int_{\Theta_{-i}} u_i(\tilde{c}(\hat{\theta}), \tilde{\mathbf{c}}(\theta_{-i}), \theta_i) dF_{-i}(\theta_{-i}) \quad (2)$$

for all $i \in N$ and $\hat{\theta} \in [0, 1]$, where $\tilde{\mathbf{c}}(\hat{\theta}_{-i})$ is the profile of all other emissions given that all other countries report their own types and comply with the recommended quota and F_{-i} is the joint probability distribution of θ_{-i} . That is, given the quotas associated with each announced type,

every type of every country weakly prefers to report its true type.

The second is the participation constraint, which states that each country must prefer to join the agreement under its recommended quota to leaving the agreement. This requires that the agreement must offer every type of country at least the amount of utility as it would get by not being in the agreement:

$$\int_{\Theta_{-i}} u_i(\tilde{c}(\theta_i), \tilde{c}(\theta_{-i}), \theta_i) dF_{-i}(\theta_{-i}) \geq \int_{\Theta_{-i}} \max_{c_i \geq 0} u_i(c_i, \tilde{c}(\theta_{-i}), \theta_i) dF_{-i}(\theta_{-i}) - K \quad (3)$$

for all i and θ_i . That is, each country's payoff for going it alone is its payoff-maximizing emissions level minus the non-participation penalty K . The agreement must then offer every type of every country at least this payoff.

We solve for the best mechanism that satisfies the constraints (2) and (3). The incentive compatibility condition (2) represents the idea of the Revelation Principle: if an outcome is implementable at all then it must be implementable by a direct mechanism in which all players are incentivized to reveal their true types. Thus, restricting ourselves to incentive compatible mechanisms is natural in the sense that any achievable outcome can be implemented by one satisfying this constraint. The participation constraint (3) represents the idea that no countries can be forced to sign on to an agreement. Thus, a feasible agreement must satisfy the property that every country would sign on to and comply with the agreement rather than going it alone. An agreement that nobody would sign on to clearly cannot be effective. However, this constraint is limiting in a particular way: we look for the optimal mechanism among those that are fully inclusive, that is, that *all* countries would sign on to. As we discuss in the next section and in the Appendix, with no investigations it may not be the case that the optimal mechanism overall is fully inclusive.

The IO must maximize the sum of all these utilities based on type:

$$\sum_{i=1}^n [\theta_i \log(\tilde{c}(\theta_i)) + (1 - \theta_i) \log(\omega - C)] \quad (4)$$

subject to (2) and (3). We state the full compression result:

Proposition 1. Harrison and Lagunoff (2017) *Assume $\underline{\theta} > \frac{\bar{\theta}}{n(1+\bar{\theta})+\bar{\theta}}$. The optimal mechanism under incomplete information in the baseline model is fully compressed: all countries must meet the*

same quota regardless of type announcements.

Under the incentive compatibility constraint, the IO can only assign all types of countries the same quota. Such a result is referred to as “fully compressed,” in that there is one quota for all countries despite these countries having different types. The reason this result is the same for each country is because all countries have the incentive to report a type higher than their true type to get a more lenient quota and therefore maximize their own utility. In other words, if quotas depend on countries’ private information, then each country can indirectly free ride by reporting a type higher than its true type. This agreement is inefficient compared to the optimal agreement because it treats all countries as equal even though their costs of carbon abatement differ.⁴⁰ Notably, Proposition 1 only makes use of the truth-telling constraint. We have not yet analyzed how the participation constraint affects the optimal quota.

Clearly, a fully compressed quota meets the incentive compatibility constraint because no country can have a strict incentive to lie. To find the optimal quota we can therefore find the best compressed quota that satisfies all countries’ participation constraints. This calculation boils down to finding the strictest quota that the highest type will be willing to accept. Given a compressed quota c^* the highest type’s utility if it opts out of the treaty is

$$\max_{c_H} \bar{\theta} \log(c_H) + (1 - \bar{\theta}) \log(\omega - c_H - (n - 1)c^*). \quad (5)$$

Solving this maximization problem for c_H gives $c_H(c^*) = \bar{\theta}(\omega - c^*(n - 1))$.

To get some intuition about the optimal compressed quota, it is useful to consider two extremes. The highest type incurs no cost from violating the agreement when K (the cost of not being in the agreement) is zero, and so the compressed quota c^* cannot be higher than the value at which the highest type receives an equal utility from remaining or leaving the agreement. This indifference condition is met only when the quota is set to the amount the highest type of country would choose to emit were it not in the agreement:

$$c^* = \frac{\bar{\theta}\omega}{1 + \bar{\theta}(n - 1)}. \quad (6)$$

⁴⁰Though we model a situation in which countries are symmetric ex ante for simplicity, an additional issue with compressed agreements may be exacerbating inequality, because policy and non-policy determinants of the costs of investment in clean technology are associated with national income (Baldwin et al. 2016.)

This is the upper bound on the optimal quota. In the literature, this issue of agreements' commitment levels being dragged down by the least committed country is known as the law of the least ambitious program.⁴¹ Essentially, in agreements that try to achieve broad membership but only require the same quota of all countries, it is only possible to assign every country a “shallow” quota that is low enough for even the least-committed country to meet (or else risk losing members).

If we assume that K is so large that the participation constraint no longer has an effect, then the compressed quota is simply the average quota under full information conditions:

$$\sum_{i=1}^n [\theta_i \log(c^*) + (1 - \theta_i) \log(\omega - c^* n)]. \quad (7)$$

Taking first-order conditions and solving for c^* (again using $T = \sum_{i=1}^n \theta_i$) yields $c^* = \frac{(\frac{T}{n})\omega}{n}$. Notice that this is a particularly intuitive mechanism: it simply asks everyone to follow the average quota under the full information solution. Note also that this is generally a lower quota than the one when K is zero. This illustrates the following result.

Proposition 2. *The optimal quota is decreasing in K with $\lim_{K \rightarrow 0} c^* = \frac{\bar{\theta}\omega}{1+\bar{\theta}(n-1)}$ and $\lim_{K \rightarrow \infty} c^* = \frac{(\frac{T}{n})\omega}{n}$.*

Proposition 2 gives us some insights about how an organization must design an agreement when it has no independent investigative power. Recall that the parameter K is meant to represent the IO's enforcement leverage. For instance, if the IO controls access to a club good as suggested by Nordhaus (2015) or if recognition by the IO is very valuable to the countries, then K may be large in the sense that the IO has leverage to make countries comply with an agreement. An insight from Proposition 2 is that this power is useful but not good enough to implement an optimal agreement. The reason is that this enforcement leverage enters countries' participation constraints but not their incentive compatibility constraints. In other words, an IO with a lot of leverage can induce countries to comply with the recommended quotas but cannot make them tell the truth about their types. As we show in the next section, introducing investigative powers helps the IO convert its leverage for compliance into leverage for honest reporting.

⁴¹Hovi, Ward, and Grundig (2015); Underdal (1980)

Non-inclusive mechanisms

Proposition 2 characterizes the best mechanism among those that are incentive compatible and fully inclusive. By fully inclusive, we mean that the participation constraint is satisfied for every country so the agreement is a global one. However, the optimum among all possible mechanisms need not be fully inclusive. The central mechanism at play is the well-known broad versus deep trade-off from the literature on climate agreements. Since the optimal mechanism without investigations is fully compressed, IOs without investigative capacity can only be either broad (including many members) or deep (requiring strong commitments), given that all countries will be asked to meet the same quota despite inherent differences in these countries' capabilities. In our model, this trade-off is evident by the fact that the quota for every country is primarily a function of the participation constraint of the highest type. That is, all countries are asked to do only as much as what the least environmentally friendly country is willing to do. Alternatively, a mechanism that is not fully inclusive but instead excludes some types lets signatories make deeper commitments but carries the disadvantage of leaving the countries with the highest carbon emissions completely unconstrained.

As we show by example in the Appendix, a non-inclusive mechanism may be optimal in some cases. The structure of the optimal agreement is still similar to that in Proposition 2 in the sense that the mechanism should be fully compressed among those that sign on to the agreement to ensure incentive compatibility, but leave non-signatories (all countries above some type cutoff) free to choose their emissions levels outside of the agreement. This primarily depends on how much leverage the international community has over enforcement of an agreement (represented in our model by the parameter K). If K is near zero, meaning there is very little impetus to comply with any agreement, the participation constraints of the worst countries become very salient and it is likely that fully inclusive mechanisms will not be optimal. By contrast, as K becomes large, the optimal mechanism will be fully inclusive because all participation constraints are easily met. This also explains why the optimal mechanism in Harrison and Lagunoff (2017) is fully inclusive, because the dynamic model in that paper allows for unbounded punishments which closely corresponds to our model as K goes to infinity.⁴²

⁴²See Finnemore and Jurkovich (2014) for an empirical discussion of inclusivity in international agreements.

Mechanisms with limited investigations

We now extend the benchmark model in the following way: the IO has a fixed amount R of investigative resources available and can distribute these resources however it chooses after seeing the type announcements of each country. Let $r_i(\hat{\theta}, \hat{\theta}_{-i}) \in [0, 1]$ denote the amount of investigative resources devoted to country i when country i reports that its type is $\hat{\theta}$ and the other countries report $\hat{\theta}_{-i}$.⁴³ We assume that i 's type is verified with probability $r_i(\hat{\theta}, \hat{\theta}_{-i})$. The mechanism we consider is one where, if a country's verified type is different from its reported type, that country is removed from the agreement and incurs the cost K of noncompliance but is also free to choose its carbon emissions levels from then on.

The IO's budget constraint is that $\sum_{i=1}^n r_i(\hat{\theta}, \hat{\theta}_{-i}) = R$ for any $\hat{\theta} \in [\underline{\theta}, \bar{\theta}]^n$. Clearly if $R \geq n$, then this corresponds to a complete information case where the first-best quota is implementable as long as K is large enough. We are interested in the conditions under which the first-best quota is implementable for smaller values of R and in characterizing the optimal allocation of investigative resources in this case. Proposition 3 states our result.

Proposition 3. *Assume $\underline{\theta} > \frac{\bar{\theta}}{n(1+\bar{\theta})+\bar{\theta}}$. For any $R > 0$ there exists $K^*(R)$ such that the optimal quota under complete information is implementable by a mechanism with limited investigations when $K \geq K^*(R)$.*

This result shows that the optimal quota may be implementable when the IO can conduct investigations. Furthermore, the result shows that enforcement leverage magnifies the effect of investigative resources. If K is large, then the optimal quota can be implemented with minimal investigative resources. More specifically, given *any* level of investigative resources, there exists some value of K such that the optimal rule can be implemented with that amount of resources. This illustrates the intuition given before: the effect of investigations is to convert leverage over compliance into leverage over honest reporting of types.

In Proposition 3, unlike the result in Proposition 2, the restriction to fully inclusive mechanisms is without loss of generality. For a variable agreement the quota can be tailored to each individual

⁴³Our inclusion of investigative resources places our model in line with regulation and crime deterrence models such as Becker (1968) and Mookherjee and Png (1994), where regulators must allocate investigative resources in addition to penalties. A key difference is that investigations reveal information directly about types rather than about violations themselves (which we assume are observable). Still, the case with large K but low investigative resources relates to the low-probability and high-penalty solutions in Becker.

country and type. An interpretation of this is that the ability to implement an agreement with variability solves the broad versus deep trade-off because we need not assign shallower commitments to lower types just to accommodate more difficult cases. In fact, Proposition 3 involves implementing a quota that is optimal among all possible quotas.

Note that, in contrast to many regulatory models with investigative capacity, our IO is using resources to directly investigate countries' costs of carbon abatement rather than to monitor emissions levels. So, like the model illustrating the broad versus deep trade-off in Barrett (1994), both versions of our model assume that emissions are perfectly observable. Determining the right allocation of informational capacity between monitoring compliance and learning types may be an interesting area for future research in a version of this model in which emissions levels are not always observed. However, we expect some investigative capacity must always be devoted to monitoring types. This is because, in contrast to the crime deterrence models such as Becker (1968), the IO does not directly control the magnitude of the punishment.⁴⁴

Discussion and conclusions

Initially, our model allows us to see that the ideal agreement involves an optimal quota which is increasing in countries' true types, increasing in the total available carbon stock, and decreasing in the total number of countries. However, without investigations, the best incentive compatible agreement is fully compressed, meaning all countries must meet the same quota regardless of type announcements. This is inefficient because those countries with high and low types are made to take the same actions, and thus those with low types are not contributing as much as they could while those with high types are made to take on an unsustainable burden. The fully compressed agreement also gives rise to the familiar broad versus deep trade-off; because everyone commits to the same quota, the behavior of all countries in a fully inclusive agreement is determined by whether the least environmentalist country is willing to participate.

Only in the model with institutional oversight is variability possible. In this version of the model, we allow the institution limited resources to investigate countries' claims about their types. As

⁴⁴This is related to the fact that, in Harrison and Lagunoff (2017), optimal mechanisms need not be fully compressed when transfers are allowed. However, we agree with those authors' assessment that an IO typically cannot freely choose transfers between countries.

a result of giving the institution the ability to investigate use of varying commitments, countries become increasingly likely to tell the truth about their types as the amount of the institution’s investigative resources (related to the likelihood of being caught) or the cost of not belonging to the agreement (related to the punishment from being caught) increases. This “truth-telling” equilibrium possible under institutional oversight then allows for an optimal assignment of emissions quotas, which ultimately enables the agreement to achieve both greater membership and compliance because countries are more willing to join and make deeper commitments within agreements with variability in commitments. As such, an optimal agreement supported by sufficient investigations and punishments can be both broad and deep.

Ultimately, these results suggest that an IO without the ability to investigate countries’ claims is only able to inefficiently request the same quota of all countries, regardless of type. If the cost of not remaining in the agreement is very high, then this compressed quota is set at the average quota given all countries’ types. However, if the cost of not belonging to the agreement is close to zero, then the quota must be set at the least committed country’s willingness to abate or else risk losing members. By contrast, an IO with some ability to investigate countries’ claims can incentivize countries to reveal their true types in an effort to maximize global welfare, by assigning appropriate quotas for each country either through increasing its investigative resources or the costs of not being part of the agreement. Furthermore, if countries are able to punish defectors very effectively, as is the case in the dynamic setting of Harrison and Lagunoff (2017), then the optimal agreement can be implemented with minimal investigative resources.

Our analysis focuses on simple mechanisms in which the IO elicits reports about countries’ types and then suggests emissions quotas, but the revelation principle suggests that our analysis helps us understand what is possible in other, more familiar environments. In particular, a flexible interpretation of our model could motivate several examples from the literature of variability mechanisms paired with information collection by the international community. First, as Slechten (2020) pointed out, climate negotiations are often preceded by an umbrella convention which sets up institutions to gather countries’ information related to climate change. In this environment, we can interpret the investigations in our model in a similar way to the certifications in that paper; a role of the IO in this stage of the negotiation may be to help verify the information submitted by the parties to reduce information asymmetries. Second, escape clauses in international climate agreements

introduce a type of variability in line with that emphasized in our paper by permitting parties to be subject to different requirements based on individual circumstances. These are often paired with procedures by which countries must submit verification of reasons for exceptions from usual rules. For instance, Article 8 of the London Dumping Convention permits exceptions to some articles in certain circumstances but requires contracting parties to consult with other parties and submit documentation of reasons for the exception.⁴⁵ Our model suggests that credible implementation of situation-specific variability in obligations may depend in part on these reporting procedures.

We leave several issues for future research. For example, one promising avenue for research might involve further probing the bargaining over institutions problem. As part of this effort, researchers could investigate under what conditions powerful states would form institutions (on the climate or other issues) with both investigative and enforcement abilities. Though the revelation principle suggests that *some* game should implement agreements that are outcome-equivalent to the direct mechanisms analyzed here, this could help political scientists to predict whether such institutions are likely to be adopted given a particular existing set of bargaining protocols. A second promising avenue for future research may be to consider alternative objective functions for the IO. For instance, some IOs appear to primarily seek the approval of a small set of powerful states.⁴⁶ This may limit the ability to implement good agreements to the extent that it makes the IO reluctant to investigate the powerful countries. If the IO has sufficient difficulty committing to investigate powerful countries, then it may not be possible to implement agreements that account for the private information of those countries.

⁴⁵Marcoux (2009)

⁴⁶See Copelovitch (2010) who develops and tests this theory in the case of the International Monetary Fund.

References

- Baldwin, Elizabeth, Sanya Carley, Jennifer Brass, and Lauren MacLean. 2016. "Global Renewable Electricity Policy: A Comparative Policy Analysis of Countries by Income Status." *Journal of Comparative Policy Analysis* Forthcoming (April). <https://doi.org/10.1080/13876988.2016.1166866>.
- Barrett, Scott. 1994. "Self-Enforcing International Environmental Agreements." *Oxford Economic Papers* 46. Oxford University Press: 878–94. <http://www.jstor.org/stable/2663505>.
- . 2003. *Environment and Statecraft : The Strategy of Environmental Treaty-Making: The Strategy of Environmental Treaty-Making*. OUP Oxford. <https://books.google.com/books?id=uqrey86neSIC>.
- Barrett, Scott, and Paul H Nitze. 2007. *Why Cooperate?: The Incentive to Supply Global Public Goods*. OUP Oxford. <https://books.google.com/books?id=zNYTDAAAQBAJ>.
- Becker, Gary S. 1968. "Crime and Punishment: An Economic Approach." *Journal of Political Economy* 76 (2). University of Chicago Press: 169–217. <http://www.jstor.org/stable/1830482>.
- Bernauer, Thomas. 2013. "Climate Change Politics." *Annual Review of Political Science* 16 (1): 421–48. <https://doi.org/10.1146/annurev-polisci-062011-154926>.
- Boockmann, Bernhard, and Paul W Thurner. 2006. "Flexibility Provisions in Multilateral Environmental Treaties." *International Environmental Agreements: Politics, Law and Economics* 6 (2). Springer: 113–35.
- Cirone, Alexandra E., and Johannes Urpelainen. 2013. "Trade Sanctions in International Environmental Policy: Deterring or Encouraging Free Riding?" *Conflict Management and Peace Science* 30 (4): 309–34. <https://doi.org/10.1177/0738894213491182>.
- Clapp, J., and P. Dauvergne. 2011. *Paths to a Green World, Second Edition: The Political Economy of the Global Environment*. The Mit Press. MIT Press. <https://books.google.com/books?id=Q8rxCwAAQBAJ>.

- Copelovitch, Mark S. 2010. "Master or Servant? Common Agency and the Political Economy of Imf Lending." *International Studies Quarterly* 54 (1). [International Studies Association, Wiley]: 49–77. <http://www.jstor.org/stable/40664237>.
- Csapó, Gergely, and Rudolf Müller. 2013. "Optimal Mechanism Design for the Private Supply of a Public Good." *Games and Economic Behavior* 80. Elsevier: 229–42.
- Dai, Xinyuan. 2007. *International Institutions and National Policies*. Cambridge University Press. <https://books.google.com/books?id=2DMx361lUcUC>.
- Dannenbergh, Astrid, Bodo Sturm, and Carsten Vogt. 2010. "Do Equity Preferences Matter for Climate Negotiators? An Experimental Investigation." *Environmental and Resource Economics* 47 (1). Springer: 91–109.
- Finnemore, Martha, and Michelle Jurkovich. 2014. "Getting a Seat at the Table: The Origins of Universal Participation and Modern Multilateral Conferences." *Global Governance: A Review of Multilateralism and International Organizations* 20 (3). Brill Nijhoff: 361–73.
- Gilligan, Michael J., and Leslie Johns. 2012. "Formal Models of International Institutions." *Annual Review of Political Science* 15 (1): 221–43. <https://doi.org/10.1146/annurev-polisci-043010-095828>.
- Goldstein, Judith, and Lisa L Martin. 2000. "Legalization, Trade Liberalization, and Domestic Politics: A Cautionary Note." *International Organization* 54 (3). Cambridge University Press: 603–32.
- Harrison, Rodrigo, and Roger Lagunoff. 2017. "Dynamic Mechanism Design for a Global Commons." *International Economic Review* 58 (3): 751–82. <https://doi.org/10.1111/iere.12234>.
- Harstad, Bård. 2018. "Pledge-and-Review Bargaining." CESifo Working Paper.
- Helfer, Laurence R. 2012. "Flexibility in International Agreements." In *Interdisciplinary Perspectives on International Law and International Relations: The State of the Art*, edited by Jeffrey L. Dunoff and Mark A. Pollack, 175–96. Cambridge University Press. <https://doi.org/10.1017/CBO9781139107310.010>.

- Herzing, Mathias. 2005. "Essays on Uncertainty and Escape in Trade Agreements," January.
- Hovi, Jon, Detlef F Sprinz, Håkon Sælen, and Arild Underdal. 2019. "The Club Approach: A Gateway to Effective Climate Co-Operation?" *British Journal of Political Science* 49 (3). Cambridge University Press: 1071–96.
- Hovi, Jon, Detlef F. Sprinz, and Arild Underdal. 2009. "Implementing Long-Term Climate Policy: Time Inconsistency, Domestic Politics, International Anarchy." *Global Environmental Politics* 9 (3): 20–39. <https://doi.org/10.1162/glep.2009.9.3.20>.
- Hovi, Jon, Hugh Ward, and Frank Grundig. 2015. "Hope or Despair? Formal Models of Climate Cooperation." *Environmental and Resource Economics* 62 (4). Springer: 665–88.
- Karp, Larry, and Jinhua Zhao. 2010. "International Environmental Agreements: Emissions Trade, Safety Valves and Escape Clauses." *Revue économique* 61 (1). Presses de Sciences Po: 153–82.
- Kennard, Amanda. 2020. "My Brother's Keeper: Other-Regarding Preferences and Concern for Global Climate Change." *The Review of International Organizations*. Springer, 1–32.
- Keohane, Robert O., and David G. Victor. 2011. "The Regime Complex for Climate Change." *Perspectives on Politics* 9 (1). Cambridge University Press: 7–23. <https://doi.org/10.1017/S1537592710004068>.
- Keohane, Robert O, and David G Victor. 2016. "Cooperation and Discord in Global Climate Policy." *Nature Climate Change* 6 (6). Nature Publishing Group: 570–75.
- Koremenos, Barbara. 2001. "Loosening the Ties That Bind: A Learning Model of Agreement Flexibility." *International Organization* 55 (2). Cambridge University Press: 289–325. <https://doi.org/10.1162/00208180151140586>.
- . 2005. "Contracting Around International Uncertainty." *The American Political Science Review* 99 (4). [American Political Science Association, Cambridge University Press]: 549–65. <http://www.jstor.org/stable/30038964>.
- Koremenos, Barbara, Charles Lipson, and Duncan Snidal. 2001. "The Rational Design of International Institutions." *International Organization* 55 (4). Cambridge University Press: 761–99.

- Kucik, Jeffrey, and Eric Reinhardt. 2008. "Does Flexibility Promote Cooperation? An Application to the Global Trade Regime." *International Organization* 62 (3). [MIT Press, University of Wisconsin Press, Cambridge University Press, International Organization Foundation]: 477–505. <http://www.jstor.org/stable/40071901>.
- Marcoux, Christopher. 2009. "Institutional Flexibility in the Design of Multilateral Environmental Agreements." *Conflict Management and Peace Science* 26 (2). SAGE Publications Sage UK: London, England: 209–28.
- Martin, Lisa L. 1992. "Interests, Power, and Multilateralism." *International Organization*. JSTOR, 765–92.
- McGinty, Matthew. 2007. "International Environmental Agreements Among Asymmetric Nations." *Oxford Economic Papers* 59 (1). Oxford University Press: 45–62.
- Michaelowa, Katharina, and Axel Michaelowa. 2017. "Transnational Climate Governance Initiatives: Designed for Effective Climate Change Mitigation?" *International Interactions* 43 (1). Routledge: 129–55. <https://doi.org/10.1080/03050629.2017.1256110>.
- Milgrom, Paul R., Douglass C. North, and Barry R. Weingast. 1990. "The Role of Institutions in the Revival of Trade: The Law Merchant, Private Judges, and the Champagne Fairs." *Economics & Politics* 2 (1): 1–23. <https://doi.org/10.1111/j.1468-0343.1990.tb00020.x>.
- Mitchell, Ronald B, and Patricia M Keilbach. 2001. "Situation Structure and Institutional Design: Reciprocity, Coercion, and Exchange." *International Organization*. JSTOR, 891–917.
- Mookherjee, Dilip, and I. P. L. Png. 1994. "Marginal Deterrence in Enforcement of Law." *Journal of Political Economy* 102 (5). University of Chicago Press: 1039–66. <http://www.jstor.org/stable/2138656>.
- Nordhaus, William. 2015. "Climate Clubs: Overcoming Free-Riding in International Climate Policy." *American Economic Review* 105 (4): 1339–70. <https://doi.org/10.1257/aer.15000001>.
- Ostrom, E. 1990. *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge University Press. <https://books.google.com/books?id=hHGgCgAAQBAJ>.

- Pelc, Krzysztof J. 2009. "Seeking Escape: The Use of Escape Clauses in International Trade Agreements." *International Studies Quarterly* 53 (2). [Oxford University Press, Wiley, The International Studies Association]: 349–68. <http://www.jstor.org/stable/27735100>.
- Rosendorff, B. Peter. 2005. "Stability and Rigidity: Politics and Design of the Wto's Dispute Settlement Procedure." *The American Political Science Review* 99 (3). [American Political Science Association, Cambridge University Press]: 389–400. <http://www.jstor.org/stable/30038947>.
- Rosendorff, B. Peter, and Helen V. Milner. 2001. "The Optimal Design of International Trade Institutions: Uncertainty and Escape." *International Organization* 55 (4). [MIT Press, University of Wisconsin Press, Cambridge University Press, International Organization Foundation]: 829–57. <http://www.jstor.org/stable/3078617>.
- Slechten, Aurélie. 2020. "Environmental Agreements Under Asymmetric Information." *Journal of the Association of Environmental and Resource Economists* 7 (3). The University of Chicago Press Chicago, IL: 455–81.
- Stern, Nicholas. 2007. *The Economics of Climate Change: The Stern Review*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511817434>.
- Sykes, Alan O. 1991. "Protectionism as a "Safeguard": A Positive Analysis of the Gatt "Escape Clause" with Normative Speculations." *The University of Chicago Law Review* 58 (1). University of Chicago Law Review: 255–305. <http://www.jstor.org/stable/1599904>.
- Tavoni, Alessandro, Astrid Dannenberg, Giorgos Kallis, and Andreas Löschel. 2011. "Inequality, Communication, and the Avoidance of Disastrous Climate Change in a Public Goods Game." *Proceedings of the National Academy of Sciences* 108 (29). National Academy of Sciences: 11825–9. <https://doi.org/10.1073/pnas.1102493108>.
- Underdal, A. 1980. *The Politics of International Fisheries Management: The Case of the Northeast Atlantic*. Universitetsforlaget. <https://books.google.com/books?id=kTCgAAAAMAAJ>.
- 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). Sage Publications Sage CA: Los Angeles, CA: 243–68.

Appendix

Full information benchmark

Proof of Lemma 1. The IO solves

$$\max_{(\tilde{c}(\theta_1), \dots, \tilde{c}(\theta_i), \dots, \tilde{c}(\theta_n))} \sum_{i=1}^n [\theta_i \log(\tilde{c}(\theta_i)) + (1 - \theta_i) \log(\omega - C)] . \quad (8)$$

Let $\tilde{C}(\theta) = \sum_{i=1}^n \tilde{c}(\theta_i)$ denote the total emissions under a mechanism $\tilde{c}$ and let $T = \sum_{i=1}^n \theta_i$ denote the sum of all types. This problem generates n first-order conditions:

$$\frac{\theta_i}{\tilde{c}(\theta_i)} = \frac{1 - \theta_i}{\omega - \tilde{C}(\theta)} + \sum_{k=1}^n \frac{1 - \theta_k}{\omega - \tilde{C}(\theta)} \quad (9)$$

$$\frac{\theta_i}{n - T} = \frac{\tilde{c}(\theta_i)}{\omega - \tilde{C}(\theta)} \quad (10)$$

$$\frac{\theta_i}{n - T} [\omega - \tilde{C}(\theta)] = \tilde{c}(\theta_i). \quad (11)$$

Solving for $\tilde{C}(\theta)$ gives:

$$\tilde{C}(\theta) = \sum_{i=1}^n \frac{\theta_i}{n - T} [\omega - \tilde{C}(\theta)] \quad (12)$$

$$= \frac{T}{n - \bar{\theta}} [\omega - \tilde{C}(\theta)] \quad (13)$$

$$\tilde{C}(\theta) = \frac{T\omega}{n}. \quad (14)$$

Plugging this solution into (11) gives $\tilde{c}(\theta_i) = \frac{\omega\theta_i}{n}$ as claimed. $\square$

Proof of Proposition 1

To prove this result, we follow the general steps in Harrison and Lagunoff (2017). First, we consider the IO's "relaxed problem" of maximizing welfare subject only to (2). We show that, in a solution to this problem, no type is allowed emissions more than $\frac{\bar{\theta}\omega}{n}$, the amount of emissions allowed for the highest type in the full information solution in Lemma 1. Second, we show that all types would choose emissions higher than $\frac{\bar{\theta}\omega}{n}$ if they freely chose their own emissions. This implies that the only

solutions to this relaxed problem are fully compressed because all types for all countries i would be incentivized to misrepresent their true type by reporting $\hat{\theta} > \theta_i$.

Lemma 2. *Let c^0 be a solution to the IO's relaxed problem of maximizing (4) subject to (2). Then $c^0(\hat{\theta}_i, \theta_{-i}) \leq \frac{\bar{\theta}\omega}{n}$ for all i with probability 1.*

Proof. Our proof follows the steps in Harrison and Lagunoff (2017). In line with their proof, we establish similar notation. We write the utility of a type θ_i of player i of a consumption plan c as

$$u_i(\theta_i, \theta_{-i}; c) = r(c) - \theta_i q_i(c)$$

where $r(c) = \log(\omega - C)$ and $q_i(c) = \log(\omega - C) - \log(c_i)$. This is simply a rewriting of the payoff where r represents the (common) rewards to conservation and q_i represents the individual costs. Consider a solution c^0 and assume that $c_i^0(\theta) > \bar{c} \equiv \frac{\bar{\theta}\omega}{n}$ for some individual i and type realizations θ . Denote the interim values of the cost and reward functions defined above by

$$\begin{aligned} R_i^0(\theta_i) &= \int_{\theta_{-i}} r(c^0(\theta)) dF_{-i}(\theta_{-i}) \text{ and} \\ Q_i^0 &= \int_{\theta_{-i}} q_i(c^0(\theta)) dF_{-i}(\theta_{-i}). \end{aligned}$$

Following Harrison and Lagunoff (2017) we can rewrite the relaxed problem as

$$\max_c \sum_i \left[R_i^0(\bar{\theta}) - \bar{\theta} Q_i(\bar{\theta}) + \int_{\underline{\theta}}^{\bar{\theta}} F_i(\theta) Q_i(\theta) d\theta_i \right] \quad (15)$$

subject to Q_i weakly decreasing. We construct an alternative consumption plan c^{**} as follows:

$$\begin{aligned} c_i^{**}(\theta) &= \begin{cases} \bar{c} & \text{if } \theta_i = \bar{\theta} \\ c_i^0(\theta) & \text{otherwise.} \end{cases} \\ c_j^{**}(\theta) &= \begin{cases} \bar{c} & \text{if } \theta_k = \bar{\theta} \text{ for any } k \\ c_j^0(\theta) & \text{otherwise.} \end{cases} \end{aligned}$$

We can write the difference in the objective functions for these two consumption plans as

$$\sum_i \left[R_i^{**}(\bar{\theta}) - \bar{\theta} Q_i^{**}(\bar{\theta}) + \int_{\underline{\theta}}^{\bar{\theta}} F_i(\theta) Q_i^{**}(\theta) d\theta_i \right] - \sum_i \left[R_i^0(\bar{\theta}) - \bar{\theta} Q_i^0(\bar{\theta}) + \int_{\underline{\theta}}^{\bar{\theta}} F_i(\theta) Q_i^0(\theta) d\theta_i \right] \quad (16)$$

$$= \sum_i \left[R_i^{**}(\bar{\theta}) - \bar{\theta} Q_i^{**}(\bar{\theta}) - (R_i^0(\bar{\theta}) - \bar{\theta} Q_i^0(\bar{\theta})) \right] + \sum_i \left[\int_{\underline{\theta}}^{\bar{\theta}} F_i(\theta) Q_i^{**}(\theta) d\theta_i - \int_{\underline{\theta}}^{\bar{\theta}} F_i(\theta) Q_i^0(\theta) d\theta_i \right] \quad (17)$$

The first sum in (17) must be positive because $\bar{c}$ is the unconstrained optimal consumption for a player of type $\bar{\theta}$ regardless of the type profile of the other players. The second sum is positive because $c^{**}(\theta) \leq c^0(\theta)$ and Q_i is monotone in c . Therefore, c^{**} has an overall higher value of the objective function in (15), contradicting the statement that c^0 is optimal. $\square$

Lemma 3. *If $\underline{\theta} > \frac{\bar{\theta}}{n(1+\bar{\theta})+\bar{\theta}}$ then for all types of all countries we have*

$$\arg \max_c \int_{\Theta_{-i}} \theta_i \log(c) + (1 - \theta_i) \log(\omega - c - \sum_{j \neq i} c_j^0(\theta_{-i})) dF_{-i}(\theta_{-i}) > \frac{\bar{\theta}\omega}{n}.$$

That is, if each country could freely choose its consumption, it would prefer to choose a level greater than that allowed to the highest type in the unconstrained optimum.

Proof. Country i 's first-order condition (using Leibniz's rule) is

$$\int_{\Theta_{-i}} \left[\frac{\theta_i}{c} - \frac{1 - \theta_i}{\omega - c - C_{-i}(\theta_i, \theta_{-i})} \right] dF_{-i}(\theta_{-i}) = 0. \quad (18)$$

let c^* denote a solution to this first-order condition. Note that country i 's optimal choice is always decreasing in the total amount consumed by the other players. Therefore c^* must be greater than the amount i would consume if all other players consumed $\bar{c} = \frac{\bar{\theta}\omega}{n}$ (the maximum amount possible

given Lemma 2) given any θ . We let $\hat{c}$ denote this amount and solve for it as follows:

$$\int_{\Theta_{-i}} \left[\frac{\theta_i}{\hat{c}} - \frac{1 - \theta_i}{\omega - \hat{c} - \frac{n}{1-n}\bar{\theta}\omega} \right] dF_{-i}(\theta_{-i}) = 0 \quad (19)$$

$$\frac{\theta_i}{\hat{c}} = \frac{1 - \theta_i}{\omega - \hat{c} - \frac{n}{1-n}\bar{\theta}\omega} \quad (20)$$

$$\hat{c} = \frac{\theta_i \omega}{n} (n(1 + \bar{\theta}) - \bar{\theta}). \quad (21)$$

Note that this is increasing in θ_i . Using that our assumption that $\underline{\theta} > \frac{\bar{\theta}}{n(1+\bar{\theta})+\bar{\theta}}$ we have

$$\hat{c} = \frac{\theta_i \omega}{n} (n(1 + \bar{\theta}) - \bar{\theta}) \quad (22)$$

$$\geq \frac{\bar{\theta}}{n(1 + \bar{\theta}) + \bar{\theta}} \frac{\omega}{n} (n(1 + \bar{\theta}) - \bar{\theta}) \quad (23)$$

$$= \frac{\bar{\theta}\omega}{n}. \quad (24)$$

Thus, we have $c^* > \hat{c} > \frac{\bar{\theta}\omega}{n}$ as claimed. $\square$

We are now ready to prove the main result.

Proof of Proposition 1. Suppose c^0 is a solution to the planner's problem and is not fully compressed. Then for some player i and type θ_i we have $c_i^0(\theta_i, \theta_{-i}) < c_i^0(\theta', \theta_{-i})$ for $\theta' \neq \theta_i$. By Lemma 2, we have $c_i^0(\theta_i, \theta_{-i}) < c_i^0(\theta', \theta_{-i}) < \frac{\bar{\theta}\omega}{n}$. But the interim expected utility for type θ_i of player i is strictly concave and, by Lemma 3, maximized at some value $c > \frac{\bar{\theta}\omega}{n}$. This implies that type θ_i of player i prefers any consumption level in $(c_i^0(\theta_i, \theta_{-i}), \frac{\bar{\theta}\omega}{n}]$ to $c_i^0(\theta_i, \theta_{-i})$. In particular, type θ_i of player i strictly prefers $c_i^0(\theta', \theta_{-i})$ to $c_i^0(\theta_i, \theta_{-i})$, contradicting the truth-telling constraint. Thus, any solution to the planner's problem must be fully compressed. $\square$

Limited investigations

Proof of Proposition 3. Recall that the optimal quota is $\tilde{c}(\theta_i) = \frac{\omega\theta_i}{n}$. Type interim expected utility of type θ_i of player i for participating in an agreement with the optimal quota given θ_{-i} is therefore

$$\theta_i \log \left(\frac{\omega\theta_i}{n} \right) + (1 - \theta_i) \log \left(\omega - \frac{\omega\theta_i}{n} - \frac{\omega}{n} \sum_{j \neq i} \theta_j \right) \quad (25)$$

The utility to type θ_i of player i if i does not participate is

$$\max_{c \geq 0} \theta_i \log(c) + (1 - \theta_i) \log \left(\omega - c - \frac{\omega}{n} \sum_{j \neq i} \theta_j \right) - K. \quad (26)$$

Taking first-order conditions and solving for the optimal c at each $\theta_{-i} \in \Theta_{-i}$ gives:

$$\frac{\theta_i}{c} = \frac{1 - \theta_i}{\omega - c - \frac{\omega}{n} \sum_{j \neq i} \theta_j} \quad (27)$$

$$\Rightarrow c = \theta_i \left(\omega - \frac{\omega}{n} \sum_{j \neq i} \theta_j \right). \quad (28)$$

Define

$$U_i^O(\theta_i) := \theta_i \log \left(\theta_i \left(\omega - \frac{\omega}{n} \sum_{j \neq i} \theta_j \right) \right) + (1 - \theta_i) \log \left(\omega - \theta_i \left(\omega - \frac{\omega}{n} \sum_{j \neq i} \theta_j \right) - \frac{\omega}{n} \sum_{j \neq i} \theta_j \right). \quad (29)$$

The utility to type θ_i of player i for opting out of the agreement is therefore

$$U_i^O(\theta_i) - K. \quad (30)$$

Clearly, for $K > U_i^O(\theta_i) - \theta_i \log \left(\frac{\omega\theta_i}{n} \right) - (1 - \theta_i) \log \left(\omega - \frac{\omega\theta_i}{n} - \frac{\omega}{n} \sum_{j \neq i} \theta_j \right)$, the participation constraint is met.

Next, consider the payoff to type θ_i of player i to submitting a report $\hat{\theta} \neq \theta_i$ given the investigation mechanism $r_i(\hat{\theta}_i, \hat{\theta}_{-i})$. In this case, player i gets its reservation payoff from (30) with probability $r_i(\hat{\theta}_i, \hat{\theta}_{-i})$ and, with probability $1 - r_i(\hat{\theta}_i, \hat{\theta}_{-i})$, gets its payoff from successfully imitating type $\hat{\theta}$ in

the optimal mechanism. This expected payoff is

$$\int_{\Theta_{-i}} \left[(1 - r_i(\hat{\theta}, \hat{\theta}_{-i})) \left[\theta_i \log \left(\frac{\omega \hat{\theta}}{n} \right) + (1 - \theta_i) \log \left(\omega - \frac{\omega \hat{\theta}}{n} - \frac{\omega}{n} \sum_{j \neq i} \theta_j \right) \right] + r_i(\hat{\theta}, \hat{\theta}_{-i}) [U_i^O(\theta_i) - K] \right] dF_{-i}(\theta_{-i}) \quad (31)$$

The truth-telling constraint is

$$\int_{\Theta_{-i}} \left[\theta_i \log \left(\frac{\omega \theta_i}{n} \right) + (1 - \theta_i) \log \left(\omega - \frac{\omega \theta_i}{n} - \frac{\omega}{n} \sum_{j \neq i} \theta_j \right) \right] dF_{-i}(\theta_{-i}) \geq \quad (32)$$

$$\int_{\Theta_{-i}} \left[(1 - r_i(\hat{\theta}, \hat{\theta}_{-i})) \left[\theta_i \log \left(\frac{\omega \hat{\theta}}{n} \right) + (1 - \theta_i) \log \left(\omega - \frac{\omega \hat{\theta}}{n} - \frac{\omega}{n} \sum_{j \neq i} \theta_j \right) \right] + r_i(\hat{\theta}, \hat{\theta}_{-i}) [U_i^O(\theta_i) - K] \right] dF_{-i}(\theta_{-i}).$$

A sufficient condition is to satisfy this constraint for every θ_{-i} , so we can write the constraint as

$$\theta_i \log \left(\frac{\omega \theta_i}{n} \right) + (1 - \theta_i) \log \left(\omega - \frac{\omega \theta_i}{n} - \frac{\omega}{n} \sum_{j \neq i} \theta_j \right) - \quad (33)$$

$$(1 - r_i(\hat{\theta}, \hat{\theta}_{-i})) \left[\theta_i \log \left(\frac{\omega \hat{\theta}}{n} \right) + (1 - \theta_i) \log \left(\omega - \frac{\omega \hat{\theta}}{n} - \frac{\omega}{n} \sum_{j \neq i} \theta_j \right) \right] - r_i(\hat{\theta}, \hat{\theta}_{-i}) [U_i^O(\theta_i) + K].$$

For $\hat{\theta} > \theta_i$, (33) holds with equality if

$$r_i(\hat{\theta}, \hat{\theta}_{-i}) = \frac{\left[\theta_i \log \left(\frac{\omega \hat{\theta}}{n} \right) + (1 - \theta_i) \log \left(\omega - \frac{\omega \hat{\theta}}{n} - \frac{\omega}{n} \sum_{j \neq i} \theta_j \right) \right] - \left[\theta_i \log \left(\frac{\omega \theta_i}{n} \right) + (1 - \theta_i) \log \left(\omega - \frac{\omega \theta_i}{n} - \frac{\omega}{n} \sum_{j \neq i} \theta_j \right) \right]}{K - U_i^O(\theta_i) + \left[\theta_i \log \left(\frac{\omega \hat{\theta}}{n} \right) + (1 - \theta_i) \log \left(\omega - \frac{\omega \hat{\theta}}{n} - \frac{\omega}{n} \sum_{j \neq i} \theta_j \right) \right]} \quad (34)$$

By Lemma 3, we have

$$\left[\theta_i \log \left(\frac{\omega \hat{\theta}}{n} \right) + (1 - \theta_i) \log \left(\omega - \frac{\omega \hat{\theta}}{n} - \frac{\omega}{n} \sum_{j \neq i} \theta_j \right) \right] - \left[\theta_i \log \left(\frac{\omega \theta_i}{n} \right) + (1 - \theta_i) \log \left(\omega - \frac{\omega \theta_i}{n} - \frac{\omega}{n} \sum_{j \neq i} \theta_j \right) \right] > 0$$

for $\hat{\theta} > \theta_i > \frac{\bar{\theta}}{n(1+\bar{\theta})+\bar{\theta}}$. Thus, the numerator of (34) is positive and the denominator goes to ∞ as

$K \rightarrow \infty$.

Let

$$r^*(K) = \sup_{\theta_i \in \Theta_i} \sup_{\theta_{-i} \in \Theta_{-i}} \sup_{\hat{\theta} \in \Theta_i} \frac{\left[\theta_i \log\left(\frac{\omega \hat{\theta}}{n}\right) + (1-\theta_i) \log\left(\omega - \frac{\omega \hat{\theta}}{n} - \frac{\omega}{n} \sum_{j \neq i} \theta_j\right) \right] - \left[\theta_i \log\left(\frac{\omega \theta_i}{n}\right) + (1-\theta_i) \log\left(\omega - \frac{\omega \theta_i}{n} - \frac{\omega}{n} \sum_{j \neq i} \theta_j\right) \right]}{K - U_i^O(\theta_i) + \left[\theta_i \log\left(\frac{\omega \hat{\theta}}{n}\right) + (1-\theta_i) \log\left(\omega - \frac{\omega \hat{\theta}}{n} - \frac{\omega}{n} \sum_{j \neq i} \theta_j\right) \right]}$$

denote the smallest amount of investigative resources that deters all types from submitting a false report given any distribution of other players' types as long as the participation constraint is met. For a given K large enough to satisfy the participation constraint, the constant investigation plan setting $r_i(\hat{\theta}_i, \hat{\theta}_{-i}) = r^*(K)$ for all $\hat{\theta}$ and all i implements the full information optimum. The total investigative budget is therefore $nr^*(K)$. Since $\lim_{K \rightarrow \infty} nr^*(K) = n \lim_{K \rightarrow \infty} r^*(K) = 0$, this shows that for any $R > 0$ there exists a value of K large enough to implement the full information optimal quota. $\square$

Extensions and additional results

An example of a non-inclusive optimal compressed mechanism

In the main text, we impose the constraint that all countries must participate and show that the optimal mechanism subject to that constraint is fully compressed. Here, we show by example that the optimal mechanism may not be fully inclusive. That is, the optimal mechanism may have some set of types opt out. By way of example, let $\omega = 2$, $n = 3$, and assume that the set of types is $[1/5, 2/5]$. We will consider the limiting case as $K \rightarrow 0$. The optimal fully inclusive mechanism from applying (6) is therefore a compressed mechanism in which $c^* = 8/13$. Our task is to show that this mechanism is not always ex-post optimal among all incentive compatible mechanisms. To show this, let's consider a realization of types such that $\theta_1 = \theta_2 = \frac{1}{5}$ and $\theta_3 = \frac{3}{5}$. Total utility under the optimal mechanism is

$$2 \left[\frac{1}{5} \log\left(\frac{8}{13}\right) + \frac{4}{5} \log\left(\frac{2}{13}\right) \right] + \frac{4}{5} \log\left(\frac{8}{13}\right) + \frac{1}{5} \log\left(\frac{2}{13}\right) \approx -3.95. \quad (35)$$

Consider an alternative mechanism for which only the lowest types would participate. That is, the mechanism is a fully compressed quota such that the lowest types are indifferent between participating and not participating and all other types opt-out. The quota that makes the lowest types indifferent (as $K \rightarrow 0$) is $c'' = \frac{2}{7}$ (again applying (6) but using the lowest type rather than the

highest type). All higher types opt out of this mechanism and choose $c_i = \theta_i(\omega - \sum_{j \neq i} c_j)$. In this case, with the type realizations above, the ex-post payoff from this mechanism is therefore

$$2 \left[\frac{1}{5} \log \left(\frac{2}{7} \right) + \frac{4}{5} \log \left(\frac{2}{7} \right) \right] + \frac{4}{5} \log \left(\frac{8}{7} \right) + \frac{1}{5} \log \left(\frac{2}{7} \right) \approx -2.65. \quad (36)$$

The additional calculations here are that $\frac{4}{5}(2 - 2\frac{2}{7}) = \frac{8}{7}$ and $2 - 2\frac{2}{7} - \frac{8}{7} = \frac{2}{7}$. That is, a non-inclusive compressed mechanism gives higher total welfare than the inclusive compressed mechanism.

This example may seem extreme because it relies on (a) a very strange realization of types at the endpoints of the support of the distribution and (b) a very extreme mechanism that only includes a measure zero set of types (i.e., only the very lowest type opts into the mechanism). However, this is only for convenience. Consider for instance a mechanism designed to make all types lower than $\frac{1}{2}$ accept. Applying (6) using $\theta = \frac{1}{2}$ rather than the highest type yields $c^* = \frac{1}{2}$. Under this mechanism (but under the same type realizations), the low types agree and choose $\frac{1}{2}$ while the high type opts out and chooses $c_i = \frac{4}{5}(2 - 2 * \frac{1}{2}) = \frac{4}{5}$. Therefore, the total utility is

$$2 \left[\frac{1}{5} \log \left(\frac{1}{2} \right) + \frac{4}{5} \log \left(\frac{1}{5} \right) \right] + \frac{4}{5} \log \left(\frac{4}{5} \right) + \frac{1}{5} \log \left(\frac{1}{5} \right) \approx -3.35. \quad (37)$$

Thus, this less extreme mechanism still outperforms the fully inclusive mechanism. Furthermore, because both expected welfare equations are continuous with respect to the type realizations except near the participation cutoff, this mechanism beats the fully inclusive mechanism for a positive measure of types.

This construction does depend on having a small value of K . In fact, for large values of K it must be the case that the optimal compressed mechanism is fully inclusive, because all of the participation constraints are easily satisfied. This explains why there is no difference between the optimal mechanism and the optimal fully inclusive mechanism in Harrison and Lagunoff (2017), where dynamic strategies resemble an unbounded value of K .

Endogenous K

The model in the text assumes an exogenous punishment K for noncompliance. Of course, international organizations do not typically have significant enforcement power, so here we consider how K may arise endogenously without the need for intrinsically powerful organizations. We consider two possibilities.

One way to think about K is as a short form for forward-looking strategies under dynamic mechanisms. We do not fully analyze this case here but instead point out the relationship to Harrison and Lagunoff (2017). That paper considers a dynamic model with the same basic elements that we present here but with a carbon stock that evolves over time as a function of the choices of the players. In the dynamic model, shirking is responded to with punishments that deplete the total carbon stock in future periods. Since these potential punishments are unbounded, the results of the dynamic model most closely resemble our limiting case of $K \rightarrow \infty$. In fact, under complete information any feasible payoff is implementable. In the case of complete information, the optimal compressed mechanism is implementable and in fact is very similar to our compressed mechanism in the limiting case of $K \rightarrow \infty$. Thus, if we interpret our model as a simplification of a richer dynamic model, we should take the punishments to be arbitrarily large. Note also that these punishments are self-enforcing and do not require state-like powers for the IO.

An alternative way to think of K is as something arising from other issues that states care about and that may be affected by climate negotiations. A benefit of this perspective is that we can still think about the effects of varying K , but this becomes a measure of the value of the non-environmental issue. Therefore, we consider an extension in which the punishment for noncompliance (or reward for compliance) is determined by issue bundling and is self-enforcing and requires no formal enforcement power by the IO. That is, the role of the IO is merely to help countries coordinate on a particular self-enforcing punishment strategy. Though climate action could in principle be bundled with any issue, we use trade as our leading example. This allows us to focus on a concrete and simple model and also places our model firmly in line with the climate club proposals of Nordhaus (2015) and others.⁴⁷ Following Rosendorff and Milner (2001), we take

⁴⁷Hovi et al. (2019); Keohane and Victor (2016)

the repeated prisoner's dilemma (PD) as the useful metaphor for trade cooperation.

Consider the following game. There is an infinite horizon with time indexed by $t \in \{0, 1, 2, \dots\}$. At time $t = 0$, the countries play the climate game exactly as described in the main text of the paper but without the exogenous penalty for noncooperation. In each period $t \geq 1$, the countries engage in a trade game which we represent as follows. In each period, every country sets trade barriers with every other country. Following Rosendorff and Milner (2001), we can represent the trade barrier game between each pair of countries as a prisoner's dilemma.⁴⁸ Thus, for each pair of countries $i, j \in N$, the stage game payoffs can be represented by the following:

		Country j	
		C	D
Country i	C	(b, b)	$(-d, a)$
	D	$(a, -d)$	$(0, 0)$

where $a > b > 0$ and $d > 0$. Since these payoffs apply to a particular pair, each country's total stage game payoff is its payoff from each of these interactions summed over all players. Formally, let $s_{ijt} \in \{C, D\}$ denote player i 's action toward player j in time t and let s_t denote the profile of all actions in time t . Let $v_{ijt}(s_{ijt}, s_{jit})$ be the pairwise stage game payoff defined in the table above (i.e., $v_{ijt}(C, C) = b, v_{ijt}(D, C) = a, v_{ijt}(D, D) = 0, v_{ijt}(C, D) = -d$). Country i 's stage game payoff is then $v_{it}(s_t) = \sum_{j \neq i} v_{ijt}(s_{ijt}, s_{jit})$. The discounted present value of a stream of payoffs is therefore $\sum_{t=1}^{\infty} \delta^{t-1} v_{it}(s_t)$ where $\delta \in (0, 1)$ is a common discount factor.

There are a great many equilibria to this game. Our purpose is merely to show that (a) punishing climate defectors can be self-enforcing in the trade stage of this game and (b) anticipation of the trade stage can endogenously generate the equivalent of our K without any formal enforcement power by the IO. We begin by analyzing the trade stage of the game. Suppose that a (possibly empty) subset of countries $M \subseteq N$ are considered non-compliant with the climate agreement and the rest are compliant. We set up a strategy profile in which compliant countries play grim trigger strategies with one another and always defect on countries in M , and where countries in M always

⁴⁸The prisoner's dilemma setup is derived from a continuous game in Rosendorff and Milner (2001) by setting the payoff when both countries defect equal to the countries' payoffs from the static Nash equilibrium to the game, the payoff when both players cooperate equal to the payoffs from the Pareto optimal pair of trade barriers, and the payoff when one player defects and the other cooperates equal to the payoff obtained when one player acts according to the Pareto optimal trade barriers and the other chooses the optimal defection given that player's action.

defect regardless of which countries they are matched with. To be complete, the strategy is described as follows for all countries $i \in N$:

- If $i \in M$ then $s_{ijt} = D$ for all $j \in N \setminus \{i\}$ and all t .
- If $i \notin M$ then $s_{ijt} = D$ for all $j \in M$ and $s_{ij1} = C$ for all $j \notin M$, and for $t > 1$ $s_{ijt} = C$ if and only if $s_{ijt'} = s_{jit'} = C$ for all $t' < t$, otherwise $s_{ijt} = D$.

No $i \in M$ will deviate from this strategy: defection is a dominant strategy in the static game and cooperating in any period cannot change future payoffs, so defection remains a best response for $i \in M$. Similarly, because countries in M will never cooperate, no country $i \notin M$ will deviate from the part of its strategy that dictates $s_{ijt} = D$ for all $j \in M$. Finally, in pairs where neither country is in M , neither country has any incentive to deviate from the grim trigger punishment strategy for the same reasons as above. Thus, this strategy profile is an equilibrium as long as pairs of climate-compliant countries have enough incentive to cooperate along the path of play. Since actions are chosen independently across pairs (e.g., player i defecting on player j affects future rounds with player j but does not affect pairs involving player i but not player j), we can simply analyze the standard PD game within each pair. Player i 's expected payoff in interactions with player j from cooperating is

$$b + \delta b + \delta^2 b + \dots = \frac{b}{1 - \delta}.$$

Player i 's expected payoff in interactions with player j from defecting is

$$a + \delta 0 + \delta^2 0 + \dots = a.$$

Therefore, player i cooperates along the path of play if $\frac{b}{1 - \delta} \geq a$ which holds when $\delta \geq \frac{a - b}{a}$. Thus, this strategy profile is an equilibrium when $\delta \geq \frac{a - b}{a}$.

The above analysis makes a very simple point that is only slightly modified from standard analyses of the repeated PD. However, we now consider the implications for the climate agreement at time $t = 0$. We consider whether all countries will choose to participate in a particular agreement. Consider an agreement that prescribes a profile of carbon consumption levels $(\tilde{c}_1(\theta_1), \dots, \tilde{c}_n(\theta_n))$.

We consider a profile in which all countries comply and ask whether any particular country would unilaterally withdraw. We first consider the fully compressed mechanism, so we drop dependence of the quota on θ_i for the moment and ignore truth-telling constraints. The utility of a country of type θ for complying given the trade equilibrium above is:

$$u_i(\tilde{c}, \tilde{c}_{-i}, \theta) + \delta \frac{b(n-1)}{1-\delta}.$$

That is, a country's expected utility for complying is now the static utility from carbon consumption plus the discounted trade benefits from being in compliance. The total trade benefits are the pairwise benefit from above multiplied by the number of other players, because the player contemplating a deviation expects all other players to remain in compliance.

The expected benefit of noncompliance is

$$\max_{c \geq 0} u_i(c, \tilde{c}_{-i}, \theta).$$

That is, a country that does not comply gets the benefit of optimally choosing its carbon consumption given the levels of the other players, but it expects zero trade benefits in the future because every country will defect in every interaction. Therefore, the country complies if

$$u_i(\tilde{c}, \tilde{c}_{-i}, \theta) + \delta \frac{b(n-1)}{1-\delta} \geq \max_{c \geq 0} u_i(c, \tilde{c}_{-i}, \theta)$$

or

$$u_i(\tilde{c}, \tilde{c}_{-i}, \theta) \geq \max_{c \geq 0} u_i(c, \tilde{c}_{-i}, \theta) - \delta \frac{b(n-1)}{1-\delta}.$$

Setting $K = \delta \frac{b(n-1)}{1-\delta}$, this is exactly the participation constraint from (3). This analysis holds for the fully compressed mechanism but works the same for the optimal mechanism with investigations. If a country is labelled noncompliant when investigations reveal that it submitted a false report, we simply set $K = \delta \frac{b(n-1)}{1-\delta}$ to obtain the same truth-telling constraint as in (33).

This simple exercise is intended merely to demonstrate how we can think of K in the original model as coming from a mechanism that does not depend on having an IO with formal enforcement power. Taking the PD model a bit more seriously, however, we could characterize how K changes as a

function of the underlying parameters. We see that K is larger when players are more patient, when the total number of countries negotiating increases, and when the value of preferential treatment on trade increases. As we mentioned, however, we think of trade as only one possible source of issue linkage. In addition, a more detailed model of the linked issue may yield further interesting insights about enforcement.