Information and Flexibility in the Design of International Climate Agreements

Jordan H. McAllister*and Keith Schnakenberg†

June 3, 2020

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 mitigate 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 such adjustment and can allow implementation of the socially optimal agreement. Furthermore, when the organization has significant enforcement power the optimal agreement is achievable with minimal investigation costs. The results suggest that discussions about institutions for climate cooperation should focus on information collection as well as enforcement.

*Department of Political Science, Washington University in St. Louis, jordan.hale@wustl.edu

†Department of Political Science, Washington University in St. Louis, keschnak@wustl.edu

Introduction

The international community has failed to make significant progress at slowing the pace of climate change even though its causes are known and its effects are potentially devastating (Barrett and Nitze 2007; Stern 2007). One problem to solve is how to design an international framework for climate cooperation that incentivizes countries to adopt emissions-reducing policies while distributing costs appropriately (Milgrom, North, and Weingast 1990; Kosfeld, Okada, and Riedl 2009). Encouraging compliance is difficult because countries have an incentive to free ride: 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 (Barrett and Nitze 2007; Hovi, Sprinz, and Underdal 2009; Bernauer 2013; Hovi, Ward, and Grundig 2015; Nordhaus 2015). The central goal of an international climate agreement is to overcome this free-riding problem (Ostrom 1990). Additionally, a climate agreement would ideally include flexibility to account for individual differences; since some countries are able to reduce fossil fuel emissions at a much lower cost than others, a one-size-fits-all policy is suboptimal in terms of global welfare (Bernauer 2013). Therefore, we consider how to achieve the joint goals of overcoming free riding and accounting for individual differences.

Unfortunately, as Harrison and Lagunoff (2017) demonstrate, simultaneously overcoming free riding and accounting for individual differences may be difficult. An institution designed to account for individual differences may create indirect free-riding incentives because countries are tempted to exaggerate their costs of carbon abatement in order to receive more allowances under the agreement (Herzing 2005; Helfer 2012). This problem is significant enough that in some circumstances the best agreement that can be implemented is one that contains no flexibility (Harrison and Lagunoff 2017). In this paper we consider how an international organization may overcome this problem if it has access to investigative resources that can be used to verify information about countries' costs of carbon abatement.

The international organization (IO) in our model is strong in the sense that it has some leverage for inducing limited compliance with an agreement (Harrison and Lagunoff 2017). This leverage may come from controlling access to a club good (Nordhaus 2015) or from some other benefit that accrues to countries for being in good standing with the organization (Stern 2007). Nonetheless, without any investigative resources, this leverage for inducing compliance does not help the organization

design a flexible agreement (Barrett and Nitze 2007). Even as the IO becomes extremely powerful it cannot prevent indirect free riding through exaggerating compliance costs (Barrett 2003; Barrett and Nitze 2007; Helfer 2012; Bernauer 2013; Nordhaus 2015). However, when the IO has the power to investigate the compliance costs of the countries it may be able to overcome free riding while also allowing flexibility (Dai 2007). More interestingly, the availability of investigative resources means that the IO's leverage for encouraging compliance can also help it incentivize truth telling in order to make the agreement more flexible (Pelc 2009). Thus, a powerful organization may be able to implement the optimal flexible agreement using minimal investigative resources.

We contribute to the literature in two main ways. First, we contribute to the literature on models of climate change cooperation. Most broadly, we follow the literature in viewing the main function of climate agreements as attempting to overcome problems of free riding (Ostrom 1990; Hovi, Ward, and Grundig 2015; Nordhaus 2015). The closest paper is Harrison and Lagunoff (2017) which takes 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 by the IO as a potential solution to this problem.

Second, we contribute to the international relations literature on flexibility in the design of international institutions more broadly. Much of the work on flexibility comes from the literature on trade agreements specificallly (Sykes 1991; Goldstein and Martin 2000; Rosendorff and Milner 2001; Herzing 2005; Rosendorff 2005; Pelc 2009).[1] In these papers, two factors driving the need for flexibility are changes over time that require renegotiation (Koremenos 2001; Koremenos, Lipson, and Snidal 2001; Barrett 2003; Kucik and Reinhardt 2008; Harstad 2018) and individual differences between countries leading to a need for escape clauses (Sykes 1991; Koremenos, Lipson, and Snidal 2001; Rosendorff and Milner 2001; Koremenos 2005; Helfer 2012) or compensatory payments (Herzing 2005; Kucik and Reinhardt 2008). We focus on the latter issue but note that the issues of asymmetric information raised in this paper may also apply to renegotiation of existing agreements (Boockmann

[1] See Gilligan and Johns (2012) for a useful review of some of this literature.

and Thurner 2006).

Our underlying argument is that flexibility may increase the number of willing signatories by ensuring that countries that face higher compliance costs receive some relief (McGinty 2007; Keohane and Victor 2011). In the literature on climate agreements this relates to the commonly noted "broad versus deep" trade-off between the strictness of actions required under an agreement and the ability to gain more signatories (Barrett 1994, 2003; Von Stein 2008; Bernauer 2013; Hovi, Ward, and Grundig 2015; Keohane and Victor 2016). 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 (Barrett 2003; Bernauer 2013). So theoretically, an agreement that allows flexibility 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, since these countries know they will not be asked to do more than they're able at any given time (Rosendorff and Milner 2001; Barrett 2003; Rosendorff 2005; Von Stein 2008; Gilligan and Johns 2012). Our conception of flexibility in this paper is the extent to which the agreement's requirements account for individual differences (Keohane and Victor 2011). A sufficiently flexible agreement softens the trade-off between breadth and depth but requires independent investigatory resources for the IO in order to overcome the incentive of countries to exaggerate costs of compliance (Milgrom, North, and Weingast 1990; Rosendorff and Milner 2001; Rosendorff 2005; Gilligan and Johns 2012). We now turn to the model.

Benchmark model

We will model the role of an institution in facilitating climate change cooperation. We consider an international organization at the helm of climate change mitigation efforts. In addition, we consider the role of flexibility in these climate agreements, in order to see whether flexible mechanisms can be used by international organizations to overcome the breadth-versus-depth trade-off and, as a result, increase the impact of the agreement. As such, this model will allow for countries to have different types regarding how much they value consumption versus conservation of carbon (or, equivalently, their costs of carbon abatement). Countries' information about their types, however, is private, and thus the international organization must attempt to elicit accurate information about the countries

in order to implement a flexible rule. Ultimately, the organization aims to maximize the total payoffs of all the countries and does this by collecting information through the voluntary disclosures of its members and then recommending flexible emissions quotas for each country based on that information.

The game

In the benchmark model, the players comprise N countries, where $N = \{1, ..., n\}$. 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.[2] 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 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" country could be an industrializing country which relies more heavily on cheap, polluting fuels in order to sustain its economy, whereas a "low type" country could be a more developed country that is less dependent on its manufacturing sectors and instead has a reliable technology and services base. These types are private information but are 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 (IO) 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 announcement $\hat{\theta}_i$, 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. As this IO is a utilitarian planner, it seeks to maximize the sum of all countries' payoffs.

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

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

[2] In 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.

In this expression $C = \Sigma_{i=1}^{n} c_i$ 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 (since 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 (since 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 for noncooperation. This fixed cost could be seen as the loss of any club benefits (Nordhaus 2015), alliances, or simply prestige as a result of not being in the agreement.

Overall, we analyze a particular kind of 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. Therefore, our setup encompasses a very broad set of institutions.

The form of the utility functions and the choices of the IO in the benchmark model follows 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.[3]

[3]Additionally, in a dynamic model the IO may not benefit from imposing large costs of noncompliance since 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 (since the optimal quota is increasing in θ_i), but the quotas for all countries are also influenced by how much carbon is there in the first place as well as by how many countries are a part of the agreement (since 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 flexible 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 describes the reality that each country must find it in its best interest to reveal its true type given the quotas associated with each announced type. This requires that the recommended quota be no larger than each country's type is willing/able to enact:

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

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 player at least the amount of utility they would get by not being in the agreement:

$$\int_{\Theta_{-i}} u_i(\tilde{c}(\theta_i), \tilde{\mathbf{c}}(\theta_{-i}), \theta_i) dF_{-i}(\theta_{-i}) \geq \int_{\Theta_{-i}} \underset{c_i \geq 0}{max} u_i(c_i, \tilde{\mathbf{c}}(\theta_{-i}), \theta_i) dF_{-i}(\theta_{-i}) - K \tag{3}$$

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

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

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

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 players 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 in

order 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 types higher than their true types. 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 since 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. Let $\theta_H = \underset{i \in N}{max} \theta_i$. This is the highest *realized* type out of the countries, not to be confused with the highest *possible* type $\bar{\theta}$. Given a compressed quota c^* the highest type's utility if it opts out of the treaty is

$$\underset{c_H}{max} \theta_H log(c_H) + (1 - \theta_H) log(\omega - c_H - (n-1)c^*). \tag{5}$$

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

To get some intuition about the optimal compressed quota it is useful to consider two extremes. When K (the cost of not being in the agreement) is zero, then the highest type incurs no cost from violating the agreement 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:

$$c^* = \frac{\theta_H \omega}{1 + \theta_H (n-1)}.$$

Essentially, 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" (Underdal 1980) and can be seen in action across a variety of current international agreements.

By contrast, 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:

$$\Sigma_{i=1}^{n} [\theta_i log(c^*) + (1 - \theta_i) log(\omega - c^* n)]. \tag{6}$$

Taking first-order conditions and solving for c^* (again using $T = \Sigma_{i=1}^{n} \theta_i$) 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 gets us the following result.

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

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.

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 choose to distribute these resources however it chooses after seeing the type announcements of each player. 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 $\hat{\theta}_i$. 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.

The IO's budget constraint is that $\sum_{i=1}^{n} r_i(\hat{\theta}_i, \hat{\theta}_{-i}) = R$ for any $\hat{\theta} \in [\underline{\theta}, \overline{\theta}]^n$. 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* $\underline{\theta} > \frac{\overline{\theta}}{n(1+\overline{\theta})+\overline{\theta}}$. *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 given 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.

Discussion

Initially, our model allows us to see that the first-best equilibrium 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, this result does not allow for the incentive compatibility or participation constraints. The incentive compatibility constraint dictates that each type of country must be incentivized to reveal its true type, while the participation constraint states that each country must be willing to join the agreement under its recommended quota rather than acting independently. When held to these two constraints, the second-best equilibrium is fully compressed, meaning all players must meet the same quota regardless of type announcements. This second result is less efficient than the first, since those countries with high and low types are made to pay the same cost, and thus those with low types are not contributing as much as they could while those with high types are unlikely to join the agreement in the first place unless the quota is set at level of abatement low enough such that even the countries with the highest types are still willing to join.

Only in the model with institutional oversight is the first-best equilibrium possible. In this version of the model, we allow the institution to investigate countries' claims about their types. As a result of giving the institution the ability to investigate use of flexibility provisions, 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 a more flexible assignment of emissions quotas, which ultimately enables the agreement to achieve both greater membership and compliance since countries are more willing to join and make deeper commitments within flexible agreements.

Conclusion

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 is able to incentivize countries to reveal their true types in an effort to maximize global social welfare by assigning appropriate quotas for each country either through increasing its investigative resources or the costs of not being part of the agreement. So clearly, there is a benefit to introducing flexibility mechanisms like type-dependent quotas if the institution is able to both verify its information and punish those countries that make false claims. Essentially, flexibility mechanisms can allow the IO to implement socially-optimal quotas that demand from each country the most it is willing to do, in order to maximize both abatement and membership in the agreement.

The model in this paper shows us how adding flexibility to climate change agreements can make them more effective, but only if the appropriate oversight or punishment is employed. Thus, the unique contribution of our model is the finding that a truth-telling equilibrium is only possible given high enough punishment *and* investigation mechanisms. Whether this monitoring is provided by the organization itself, by concerned NGOs and other third-party groups, or willingly by the states

being investigated, it is key that the organization is able to conduct some form of oversight. This finding offers one possible explanation for why all of the recent efforts to form flexible yet effective international climate agreements have failed; either the ability of the organization to monitor its members or the exclusive benefit from being in the agreement has not been sufficient. In other words, this model suggests that agreements with monitoring capabilities, like the Kyoto Protocol, need to increase oversight further or perhaps implement a more valuable benefit structure in order to be effective. Only in this way can an institution successfully provide flexible emissions quotas to different countries depending on their true types and in this way achieve the socially-optimal equilibrium.

References

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.

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.

Dai, Xinyuan. 2007. *International Institutions and National Policies*. Cambridge University Press. https://books.google.com/books?id=2DMx361lUcUC.

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.Editors 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, 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.

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.

Kosfeld, Michael, Akira Okada, and Arno Riedl. 2009. "Institution Formation in Public Goods

Games." *The American Economic Review* 99 (4). American Economic Association: 1335–55. http://www.jstor.org/stable/25592511.

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.

McGinty, Matthew. 2007. "International Environmental Agreements Among Asymmetric Nations." *Oxford Economic Papers* 59 (1). Oxford University Press: 45–62.

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.

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.

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.

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

$$\max_{(\tilde{c}(\theta_1),\ldots\tilde{c}(\theta_i),\ldots,\tilde{c}(\theta_n))} \sum_{i=1}^{n} \left[\theta_i \log(\tilde{c}(\theta_i)) + (1-\theta_i)\log(\omega - C)\right]. \tag{7}$$

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:

$$\frac{\theta_i}{\tilde{c}(\theta_i)} = \frac{1-\theta_i}{\omega - \tilde{C}(\theta)} + \sum_{k=1}^{n} \frac{1-\theta_k}{\omega - \tilde{C}(\theta)} \tag{8}$$

$$\frac{\theta_i}{n-T} = \frac{\tilde{c}(\theta_i)}{\omega - \tilde{C}(\theta)} \tag{9}$$

$$\frac{\theta_i}{n-T}\left[\omega - \tilde{C}(\theta)\right] = \tilde{c}(\theta_i). \tag{10}$$

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

$$\tilde{C}(\theta) = \sum_{i=1}^{n} \frac{\theta_i}{n-T}\left[\omega - \tilde{C}(\theta)\right] \tag{11}$$

$$= \frac{T}{n-\overline{\theta}}\left[\omega - \tilde{C}(\theta)\right] \tag{12}$$

$$\tilde{C}(\theta) = \frac{T\omega}{n}. \tag{13}$$

Plugging this solution into (10) gives $\tilde{c}(\theta_i) = \frac{\omega\theta_i}{n}$ as claimed. □

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 $\frac{\overline{\theta}\omega}{n}$, 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 $\frac{\overline{\theta}\omega}{n}$ if they freely chose their own emissions. This implies that the only

solutions to this relaxed problem are fully compressed since all types for all countries i would be incentivized to free ride 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 $c^0(\hat{\theta}_i, \theta_{-i}) \leq \frac{\overline{\theta}\omega}{n}$ 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 $c_i^0(\theta) > \overline{c} \equiv \frac{\overline{\theta}\omega}{n}$ for some individual i and type realizations θ. Denote the interim values of the cost and reward functions defined above by

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

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

$$\max_c \sum_i \left[R_i^0(\overline{\theta}) - \overline{\theta} Q_i(\overline{\theta}) + \int_{\underline{\theta}}^{\overline{\theta}} F_i(\theta) Q_i(\theta_i) d\theta_i \right] \tag{14}$$

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

$$c_i^{**}(\theta) = \begin{cases} \overline{c} & \text{if } \theta_i = \overline{\theta} \\ c_i^0(\theta) & \text{otherwise.} \end{cases}$$

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

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

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

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

The first sum in 16 must be positive because $\overline{c}$ is the unconstrained optimal consumption for a player of type $\overline{\theta}$ 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 (14), contradicting the statement that c^0 is optimal. □

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

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

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 Liebniz's rule) is

$$\int_{\Theta_{-i}} \left[\frac{\theta_i}{c} - \frac{1-\theta_i}{\omega - c - C_{-i}(\theta_i, \theta_{-i})} \right] dF_{-i}(\theta_{-i}) = 0. \tag{17}$$

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 $\overline{c} = \frac{\overline{\theta}\omega}{n}$ (the maximum amount possible

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

$$\int_{\Theta_{-i}} \left[\frac{\theta_i}{\hat{c}} - \frac{1-\theta_i}{\omega - \hat{c} - \frac{n}{1-n}\overline{\theta}\omega} \right] dF_{-i}(\theta_{-i}) = 0 \tag{18}$$

$$\frac{\theta_i}{\hat{c}} = \frac{1-\theta_i}{\omega - \hat{c} - \frac{n}{1-n}\overline{\theta}\omega} \tag{19}$$

$$\hat{c} = \frac{\theta_i \omega}{n}(n(1+\overline{\theta}) - \overline{\theta}). \tag{20}$$

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

$$\hat{c} = \frac{\theta_i \omega}{n}(n(1+\overline{\theta}) - \overline{\theta}) \tag{21}$$

$$\geq \frac{\overline{\theta}}{n(1+\overline{\theta})+\overline{\theta}} \frac{\omega}{n}(n(1+\overline{\theta}) - \overline{\theta}) \tag{22}$$

$$= \frac{\overline{\theta}\omega}{n}. \tag{23}$$

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

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 $c_i^0(\theta_i, \theta_{-i}) < c_i^0(\theta', \theta_{-i}) < \frac{\overline{\theta}\omega}{n}$. But the interim expected utility for type θ_i of player i is strictly concave and, by Lemma 3, maximized at some value $c > \frac{\overline{\theta}\omega}{n}$. This implies that type θ_i of player i prefers any consumption level in $(c_i^0(\theta_i, \theta_{-i}), \frac{\overline{\theta}\omega}{n}]$ 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. □

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

$$\theta_i \log\left(\frac{\omega\theta_i}{n}\right) + (1-\theta_i)\log\left(\omega - \frac{\omega\theta_i}{n} - \frac{\omega}{n}\sum_{j\neq i}\theta_j\right) \tag{24}$$

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

$$\max_{c\geq 0} \theta_i \log(c) + (1-\theta_i)\log\left(\omega - c - \frac{\omega}{n}\sum_{j\neq i}\theta_j\right) - K. \tag{25}$$

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

$$\frac{\theta_i}{c} = \frac{1-\theta_i}{\omega - c - \frac{\omega}{n}\sum_{j\neq i}\theta_j} \tag{26}$$

$$\Rightarrow c = \theta_i\left(\omega - \frac{\omega}{n}\sum_{j\neq i}\theta_j\right). \tag{27}$$

Define

$$U_i^O(\theta_i) \coloneqq \theta_i \log\left(\theta_i\left(\omega - \frac{\omega}{n}\sum_{j\neq i}\theta_j\right)\right) + (1-\theta_i)\log\left(\omega - \theta_i\left(\omega - \frac{\omega}{n}\sum_{j\neq i}\theta_j\right) - \frac{\omega}{n}\sum_{j\neq i}\theta_j\right). \tag{28}$$

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

$$U_i^O(\theta_i) - K. \tag{29}$$

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 (29) 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

$$\int_{\Theta_{-i}} \left[(1 - r_i(\hat{\theta}, \hat{\theta}_{-i}) \left[\theta_i \log \left(\frac{\omega \hat{\theta}}{n} \right) + (1 - \theta_i) \log \left(\omega - \frac{\omega \hat{\theta}}{n} - \frac{\omega}{n} \sum_{j \neq i} \theta_j \right) \right] + r_i(\hat{\theta}, \hat{\theta}_{-i}) \left[U_i^O(\theta_i) - K \right] \right] dF_{-i}(\theta_{-i}) \tag{30}$$

The truth-telling constraint is

$$\int_{\Theta_{-i}} \left[\theta_i \log \left(\frac{\omega \theta_i}{n} \right) + (1 - \theta_i) \log \left(\omega - \frac{\omega \theta_i}{n} - \frac{\omega}{n} \sum_{j \neq i} \theta_j \right) \right] dF_{-i}(\theta_{-i}) \geq \tag{31}$$

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

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

$$\theta_i \log \left(\frac{\omega \theta_i}{n} \right) + (1 - \theta_i) \log \left(\omega - \frac{\omega \theta_i}{n} - \frac{\omega}{n} \sum_{j \neq i} \theta_j \right) - \tag{32}$$

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

For $\hat{\theta} > \theta_i$ (32) holds with equality if

$$r_i(\hat{\theta}, \hat{\theta}_{-i}) = \frac{\left[\theta_i \log \left(\frac{\omega \hat{\theta}}{n} \right) + (1 - \theta_i) \log \left(\omega - \frac{\omega \hat{\theta}}{n} - \frac{\omega}{n} \sum_{j \neq i} \theta_j \right) \right] - \left[\theta_i \log \left(\frac{\omega \theta_i}{n} \right) + (1 - \theta_i) \log \left(\omega - \frac{\omega \theta_i}{n} - \frac{\omega}{n} \sum_{j \neq i} \theta_j \right) \right]}{K - U_i^O(\theta_i) + \left[\theta_i \log \left(\frac{\omega \hat{\theta}}{n} \right) + (1 - \theta_i) \log \left(\omega - \frac{\omega \hat{\theta}}{n} - \frac{\omega}{n} \sum_{j \neq i} \theta_j \right) \right]} \tag{33}$$

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 $\hat{\theta} > \theta_i > \frac{\overline{\theta}}{n(1+\overline{\theta})+\overline{\theta}}$. Thus, the numerator of (33) is positive and the denominator goes to ∞ as $K \to \infty$.

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 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 $\lim_{K\to\infty} nr^*(K) = n\lim_{K\to\infty} r^*(K) = 0$ this shows that for any $R > 0$ there exists a value of K large enough to implement the full information optimal quota. □