Stern, Lennart

Conference Paper

Rewarding countries for taxing fossil fuel combustion-optimal mechanisms under exogenous budgets

Beiträge zur Jahrestagung des Vereins für Socialpolitik 2024: Upcoming Labor Market Challenges

Provided in Cooperation with:
Verein für Socialpolitik / German Economic Association

Rewarding countries for taxing fossil fuel combustion: optimal mechanisms under exogenous budgets

January 29, 2024

Abstract

Global environmental institutions have historically rewarded *projects* to reduce carbon emissions that they assess as unprofitable without their support. In contrast, this paper compares this approach to a simple alternative based on *policies*, where new global institutions would each year reward countries based on their current tax rate on the combustion of a particular fossil fuel. I develop a model in which countries differ in the co-benefits that they derive from reducing fossil fuel use and in their aversion to taxing it. I calibrate the model based on the distribution of fossil fuel tax rates observed in 2017. For coal, I find that the policy-based approach outperforms the project-based approach as soon as the annual budget at the global institution's disposal exceeds \$15billion. In a dynamic extension of the model the time of the global institution's creation is uncertain. Ex post, it is optimal for the global institution to reward countries on the basis of their changes in tax rates relative to the rates before its creation. Such a mechanism creates perverse incentives during the time before the global institution's creation. To avoid these perverse incentives, it might be beneficial for the world to adopt a norm requiring newly created global institutions to abstain from conditioning their reward payments on countries' past actions.

1 Introduction

Currently, global environmental institutions spend around \$5 billion per year on supporting carbon emission reductions in developing countries. This has so far been done predominantly through project-based approaches, where the global institutions co-finance emission-abating projects that they assess would otherwise not get implemented. The industrialized countries have pledged to increase funding to global environmental institutions to \$100 billion per year[1]. This paper argues that if the world were to raise additional funding of this magnitude then it might be optimal to spend this additional money via a new kind of global institution that would reward countries for taxing the combustion of specific fossil fuels. For brevity, I will call these institutions "carbon pricing reward funds".

A simple version of this proposal goes as follows: Separate funds are created for rewarding countries for taxing coal combustion and oil combustion, respectively. Each year, each country receives a non-negative reward payment that is proportional to a function of its tax rate on the fossil fuel. The function transforming tax rates into reward payments is the same for all countries. For this reason, I refer to these mechanisms as "anonymous" mechanisms.

I develop a simple model in which I compute the optimal anonymous mechanism. There is a continuum of countries choosing their net tax on the fossil fuel (where fossil fuel subsidies are viewed as negative taxes). Each country's utility is quasilinear in money and has three further parts.

[1]There is some ambiguity about the specific meaning of this pledge, made at the Copenhagen conference in 2010. However, Roberts et al. (2021) document: "The funds promised in Copenhagen were expected by developing countries to be dominated by public grants directed through the then new UNFCCC Green Climate Fund".

Firstly, there is the purely economic benefit derived from the use of the fossil fuel. It is to be interpreted as the sum of producer and consumer surplus due to the use of the fossil fuel . This part is assumed to be public information (i.e. known by the global institution), given that changes in the economic benefit from fossil fuel use can actually be inferred from how fossil fuel use reacts to price changes. The actual fossil fuel used in the country is determined by the tax rate on the fossil fuel as all economic actors take the tax rate as given and optimize accordingly.

Secondly, there is a privately known local co-benefit from reducing fossil fuel use. This is interpreted to include the countries' subjective valuation of the benefits of health improvements from reduced local air pollution. This term can also capture terms-of-trade effects, where countries take into account the effect of their tax rates on the world market price of the fossil fuel.

Thirdly, there is a privately known subjective aversion cost from taxing the fossil fuel (or to not subsidizing it), interpreted to arise from the popular opposition to high fuel prices and the resulting societal disruptions.

I specialize to the case where the purely economic benefits from fossil fuel use are a second order polynomial of the fossil fuel use and proportional in all countries, the local co-benefits are linear in fossil fuel use and the subjective aversion costs from fossil fuel taxation are linear in the tax rate.

In this case, the two privately known parameters can actually be reduced to a single linear combination which can be viewed as the type of the country. This type also equals the tax rate that the country chooses in the absence of the global institution. Based on this, I very roughly calibrate the model using a uniform distribution of types with ranges set to equal those of the tax rates observed in 2017 in developing countries.

The global institution knows this type distribution. It is restricted to use anonymous mechanisms. Formally, this corresponds to contracting under asymmetric information: The global institution cannot observe countries' individual types and can only condition reward payments on their chosen tax rates.

The global institution's objective is the sum of the social cost of carbon multiplied by aggregate emission reductions and the countries' current utilities. I find that the optimal anonymous mechanism changes very little if the global institution instead simply aims to maximize aggregate emission reductions, as long as its annual budget does not greatly exceed \$100billion. For improved tractability, I therefore assume for the rest of the paper that the global institution's objective is simply to minimize emissions.

Within the same model, I formalize an optimistic view of project-based contracting, the traditional approach taken by global environmental institutions. I assume that the global institution remunerates projects (interpreted to be e.g. installations of wind farms or energy efficiency programs) at a constant rate per averted emission. Optimistically, I assume that it can perfectly identify which projects would counterfactually not happen without its support.

Despite these optimistic assumptions, I find for the case of coal that the project-based approach is outperformed by the optimal anonymous mechanism as soon as the annual available budget exceeds \$15billion.

The question arises as to whether alternative mechanisms could improve upon the optimal anonymous mechanism. Specifically: Should the global institution condition its reward payments on the tax rates that countries chose before its creation? Since past tax rates contain valuable information about the countries' current preferences (i.e. their types), this could allow the global institution to get closer to the performance achievable under complete information. The static model provides an upper bound on the performance gain that could be realized by allowing the global institution to condition reward payments on variables other than the current tax rates. This bound is obtained by comparing the emission reduction achieved at the optimal anonymous mechanism to that achievable under contracting with complete information. For annual budgets of at least \$15 billion, the optimal anonymous mechanism achieves more than 70% of the emission reductions attained at the optimal mechanism under complete information in the case of coal and more than 41% in the case of oil. Thus the restriction to using anonymous mechanisms imposes a welfare loss of at most 30% in the case of coal and at most 59% in the case of oil, as long as at least \$15 billion is available per year for each of the corresponding institutions.

Having obtained this bound on the downsides of restricting the global institution to condition only on countries' current tax rates, I then explore its potential upsides in a dynamic extension of the model: it avoids potentially large perverse incentive effects that arise if the global institutions can condition reward payments on past tax rates.

In the dynamic model, there is a constant exogenous probability each year that the global institution is created. Once it is created then it rewards each country from then inwards on the basis of both its current tax rate and its tax rate just before the global institution's creation. Each year, there is a constant exogenous probability that the global institution gets dissolved. During each year of its existence, it has a constant exogenous budget that it is required to spend in that year.

The ex-post optimal mechanism involves rewarding each country purely on the basis of the increase in its tax rate. This kind of mechanism creates perverse incentives in the time before the global institution's creation: Each country knows that the lower it chooses its tax rate, the higher its future reward payments will be in case the global institution is created at that moment. Illustrative calibrations yield that the perverse incentive effect reduces the overall welfare gains that the mechanism achieves by 62.5%. This welfare loss is larger than the above-mentioned upper bounds on the welfare losses that arise if the global institution is restricted to only condition its reward payments on current tax rates, assuming that the global institutions' annual budgets exceed \$15billion. This suggests that it might be beneficial for countries to adopt a norm requiring newly created global institutions tasked with rewarding countries for globally beneficial policies to abstain from conditioning their reward payments on countries' past actions and instead aim to implement the optimal mechanism conditioning reward payments only on the current policy choices.

1.1 Related literature

Martimort and Sand-Zantman (2013,2016) develop a model for international agreements that contract on countries' emission reductions, such as the Kyoto protocol. Countries differ in a multiplicative parameter that captures their heterogeneous emission abatement costs. The authors emphasize that these costs "should be understood in a broad sense as including not only technological costs but also the opportunity and political costs of reaching a given effort target". In the current paper, I decompose these effects into the purely economic part and a component intended to capture an aversion to fossil fuel taxes.

Concerning the purely economic part, I assume homogeneity across countries. This assumption would likely be problematic when analyzing economy wide agreements like the Kyoto protocol. Countries differ in the sectoral composition of their carbon emissions. Sectors differ in their price elasticity of demand for fossil fuel. Thus aggregate price elasticity of demand for fossil fuel will differ across countries. However, in the current paper I study the design of a global institution rewarding countries for pricing a specific fossil fuel (e.g. coal for electricity generation or coal for steel production). Thus the homogeneity assumption is likely to be less problematic.

Instead of contracting on emission reductions, the global institutions analyzed in the current paper contract on the taxes that the countries impose on specific fossil fuels. Cramton & Stoft (2012,2017) argue that carbon prices have important advantages over emission quantities as variables to contract on. Carbon prices give a good measure of the effort that a country exerts for reducing emissions. Changes in emission are a less effective proxy since they are greatly influenced by other factors such as economic growth.

In line with these arguments, Steckel et al. (2017) argue that global institutions tasked with incentivizing developing countries for climate change mitigation should condition their reward payments on countries' carbon prices. They contrast such policy-based approaches with the traditional project-based approaches used by the existing global environmental institutions and argue in favor of the former. Moreover, they propose that reward payments be tailored to individual countries. Such an approach is also proposed and analyzed by Strand (2020a,2020b).

In section 6 I argue that the anonymous kind of mechanism studied in this paper potentially has some long term advantages over the approach of tailoring reward payments to individual countries. These have to be weighed against

the efficiency gains from the latter that I quantify in sections 3 and 5.2. In section 5.2 I find that even if a global institution restricts itself to only condition the reward payments on countries' carbon prices (and their counterfactual emissions in the absence of any government intervention), it will likely achieve larger emission reductions than the project-based approaches, at least for coal and for large annual budgets.

The dynamic model in section 6 contributes to the literature on the "ratchet effect". In the context of the current paper, the ratchet effect means that the global institution has ex post incentives to offer more demanding reward payment schemes to countries with high past tax rates. Ex ante, these reward payment schemes create perverse incentives for countries to lower their tax rates. Freixas, Guesnerie, & Tirole (1985) study this ratchet effect in a two-type model of which Laffont Tirole (1993, chapter 9) also consider a version with a continuum of types. In these models, the agents' types do not change over time which turns out to imply that there cannot be full separation of types in the first period (see proposition 9.1 in Laffont Tirole (1993, chapter 9)). I instead focus on the case where countries' types change sufficiently strongly for there to remain substantial rents for the agents even after the global institution conditions countries' reward payments on their past tax rates. In this case, I find that there is a time consistent mechanism such that there is full separation between countries before the global institution's creation. The illustrative calibration in section 6.2.2 suggests that this case plausibly has at least some empirical relevance. The other case, when countries' types change little over time, is beyond the scope of this paper. The results from Laffont Tirole (1993, chapter 9) indicate that there are substantial technical difficulties to be overcome in order to understand this case fully.

It is well-known that in a dynamic model consisting of a repetition of a static model if the principal and the agent have common discount rates and the agent's types do not change over time, then the optimal dynamic mechanism is simply a repetition of the optimal static mechanism (Laffont & Martimort (2009), chapter 8). In particular, it is optimal for the principal to commit to never condition reward payments on the agent's past actions. However, this result does not hold in the dynamic model that I study in section 6. There are two features of the model that set it apart from the model studied in Laffont & Martimort (2009). Firstly, the global institution (i.e the principal) has an exogenous flow of funding that it needs to spend immediately as it arrives. Secondly, its discount rate is lower than that of the agents (i.e. the countries). This lower discount rate is based on the premise that from a normative perspective the appropriate pure rate of time preference is 0 (see Greaves (2017) for a review of the arguments for and against this claim). As a result of this lower discount rate, the question as to whether or not it is good for the global institution to commit to not condition its reward payments on countries' past actions becomes an empirical one. The illustrative calibration in section 6.2.1 combined with the results from the static model (section 5.2) tentatively suggest that the answer is yes in the specific case of the carbon pricing reward funds proposed here.

1.2 Roadmap

In section 2 I start by laying out the basic static model. I then formally describe the optimal mechanisms under complete (section 2.1) and incomplete (2.2) information. Section 2.3 gives a closed form solution for the optimal mechanism under incomplete information in the limiting case where the global institution's objective is simply to minimize aggregate global emissions. In section 3.1.1 I provide an empirical calibration of the model for the cases of coal and oil. Then I show the numerical results for the optimal mechanisms for coal (3.1.2) and oil (3.2.2). Section 3.3 shows that under incomplete information the mechanism that maximizes global aggregate emission reductions differs very little from the optimal mechanism. Section 4 defines and then motivates an operational proposal for implementing carbon pricing reward funds in practice. Section 5 compares carbon pricing reward funds with the alternative approach of project-based contracting. Section 6 defines a dynamic extension of the model that I use to quantify the perverse incentive effects that could arise if the global institution is not barred from conditioning reward payments on countries' past tax rates.

2 The model

There is a continuum of countries, indexed by their type, $(\theta, \phi) \in R^2$. These two dimensions will end up reducing to only 1 relevant dimension, as I will explain further below. Each country chooses a domestic carbon tax rate τ. All the actors in the country's economy then face this carbon price and optimist accordingly. This determines the aggregate fossil fuel use $y(\tau)$ in the economy, as I will now detail.

$B(y)$ denotes the market surplus from fossil fuel use. As a result of facing the carbon price τ in addition to the exogenous[2] world market price p of fossil fuels, the actors in the economy end up jointly maximizing the following:

$$B(y) - py - \tau y$$

Let us denote by $y(\tau) := \max_\tau B(y) - py - \tau y$ the resulting fossil fuel use. It is characterized through:

$$B'(y(\tau)) = p + \tau$$

For tractability, I assume a quadratic form for the economic benefit from using fossil fuels:

Assumption 1. *The benefit from using fossil fuels is given by:*
$B(y) = \frac{y(2e_d - y + 2)}{2e_d}$

Lemma 1. *Fossil fuel demand is given by* $y(\tau) = 1 + e_d(1 - p - \tau)$.

I will later choose the normalization that $p = 1$, yielding $y(\tau) = 1 - e_d\tau$. It follows that e_d is the price elasticity of demand for fossil fuels when $\tau = 0$. I assume that this has the same value in all countries.

The country's utility is given by:

$$u(\tau, \theta, \phi) = B(y(\tau)) - py(\tau) - \theta y(\tau) + \sigma y(\tau) - \phi\tau$$

θ captures the "internalized externalities". This includes the country's valuation of local externalities from fossil fuel use and its partial internalization of climate change damages. I have omitted in the utility the effect of the other countries' emissions on the country's utility. This is justified because for now I am concerned with modeling how the country will react to incentives. I will assume that each country takes the other countries' tax rates as given and maximizes its overall utility. The country's overall utility is u plus a monetary transfer that it will receive from the global institution that I will introduce later on. The global institution will of course take global climate change damages into account, as I will explain later.

σ captures a "terms of trade effect" which arises as follows: Every unit of fossil fuel used domestically means a unit less on the world market. This increase the world market price of the fossil fuel. In fact, given our assumption of linear global demand for the fossil fuel, this decrease in the world market price is roughly a linear function of the fossil use y in the country in question, at least if the global supply for fossil fuel is approximately linear in the relevant range.

Therefore, an oil exporter's opportunity cost of using a barrel of oil is less than the world market price, since if it were to put the barrel of oil on the world market instead of using it, then this would yield additional revenue equal to the world market price, but it would also reduce the world market price of oil and thereby reduce the revenue earned on all the other barrels that it exports. From the self-interested perspective of the oil exporter, it is optimal to have the price paid by domestic users of oil to be equal to the opportunity cost of oil. Oil subsidies can

[2]This exogeneity assumption is not restrictive. In fact the current model can be embedded in a global Walrasian equilibrium model developed in Stern (2023). The analysis below can be viewed as finding out how the global institution can achieve a given reduction in fossil fuel *demand* at a minimal budget. An analogous carbon pricing reward fund could also be defined for the supply side, where a global institution could reward countries on the basis of the rates of the taxes that they levy on the extraction of coal and oil.

therefore be optimal from the self-interested perspective of the oil exporter. This explanation is supported by the data shown in section 3: Some of the largest oil exporters (Saudi-Arabia, Iran, Venezuela) have the largest subsidies on gasoline and diesel.

ϕ is meant to capture the country's aversion to taxing fossil fuel emissions. In contrast to the other terms in the country's utility function, this term is non-standard and therefore merits detailed discussion.

The first rationale for introducing the tax aversion term $-\phi\tau$ in the utility function is that some such term has to be added to the standard utility in order to account for the large empirically observed fossil fuel subsidies (see section 3.1.1 for some data). Making this term only depend on the tax rate τ and not on the economic response $y(\tau)$ to the tax rate is motivated by the following observations: Most explanations for the existence of fossil fuel subsidies are based on the distributional effects, both actual (Sterner et al. (2012), Strand (2020)) and perceived (Douenne and Fabre (2020)). These effects are primarily determined by the tax rate τ and do not greatly depend on economy's response, $y(\tau)$, to changes in the tax rate.[3]

From now on we will reason about a country's utility relative to what it gets if it chooses its tax rate to maximize its utility (not including transfers from the global institution that I will introduce below):

Definition 1.

$$\tilde{u}(\tau, \theta, \sigma, \phi) := u(\tau, \theta, \sigma, \phi) - sup_{\tau'} u(\tau', \theta, \sigma, \phi)$$

Lemma 2. *Let $\gamma := \theta - \sigma - \frac{\phi}{e_d}$. We have:*

$$\tilde{u}(\tau, \theta, \sigma, \phi) = -\frac{1}{2} e_d (\tau - \gamma)^2$$

We can thus view γ as the country's type. We denote by $v(\tau, \gamma)$ the change in utility that the type γ experiences when choosing the tax rate τ instead of choosing the tax rate that maximizes u. We thus write:

$$v(\tau, \gamma) := -\frac{1}{2} e_d (\tau - \gamma)^2$$

Since we are assuming that e_d is identical across countries, $\gamma := \theta - \sigma - \frac{\phi}{e_d}$ is a characteristic of the country that captures all that is relevant for the country's utility. We will therefore from now on view γ as being the country's type.

Here is an intuitive explanation of why $\gamma := \theta - \sigma - \frac{\phi}{e_d}$ is the country's preferred tax rate: θ is the country's valuation of the externalities from the fossil fuel. If $\sigma = 0, \phi = 0$, then clearly the standard Pigouvian logic implies that the country's preferred tax rate is θ. This of course holds generally, independently of the specification of B. Now for determining the country's preferred tax rates, there are two other effects to be taken into account: Firstly, the intrinsic aversion to taxing the fossil fuel, captured by the term $-\phi\tau$, weighs in favor of reducing the tax on the fossil fuel below θ. However, this consideration has to be weighed against a further effect: the standard deadweight loss due to distorting the quantity y. This latter effect is proportional to the elasticity parameter e_d. Thus it makes sense that the adjustment due to the tax aversion effect should be dampened in proportion to the value of e_d, which is indeed what happens in the term $-\frac{\phi}{e_d}$.

Now I introduce the global institution. This global institution has an exogenous budget M. It can observe the carbon price τ and offer countries non-negative transfers, $t(\tau)$. The country's overall utility is thus:

$$v(\tau, \gamma) + t(\tau)$$

The global institution knows the function $B(y)$ and in particular the parameter e_d (which equals the price elasticity of demand of the fossil fuel in the absence of any tax, given that we will normalize the world market price

[3]For small tax rates, this claim is true by the Envelope Theorem.

p of fossil fuels to 1.). The motivation for this assumption is that in practice the shape of $B(y)$ can be inferred from observed market responses to changes in fossil fuel prices.

I assume that the global institution cannot condition its transfer on γ. Instead, it can only condition its transfer on τ. I discuss and motivate this assumption in section 6. The global institution's objective is:

$$\int_{\gamma=\underline{\gamma}}^{\bar{\gamma}} (v(\tau(\gamma),\gamma) - \eta y(\tau(\gamma)))f(\gamma)d\gamma$$

where $f(\gamma)$ denotes the density of the type distribution and η the (global) social cost of carbon[4], as evaluated by the global institution.

The global institution's objective thus consists of two terms: Firstly, the sum of the countries' private utilities and secondly, the global external effects of fossil fuel combustion. Concerning the latter, I assume for the numerical calibrations in section 3 that the global institution uses a normative approach for determining the social cost of carbon η. This is clearly appropriate: The global institution should aim to maximize impartial global welfare, it should not base its evaluation of climate change damages on countries' revealed preferences. (In particularly, it should not use pure time discounting in evaluating future climate change damages.)[5]

The global institution's budget constraints means that its aggregate transfers cannot exceed its exogenous budget M:

$$\int_{\gamma=\underline{\gamma}}^{\bar{\gamma}} t(\tau(\gamma))f(\gamma)d\gamma \leq M$$

[4]Here I do not account for the fact that the world market will adjust as a result of the decrease in the demand. To take this into account, I should multiply $y(\tau)$ by $\frac{1}{1-\frac{\epsilon_d}{\epsilon_s}}$, where ϵ_d denotes the global price elasticity of demand and ϵ_s the global price elasticity of supply. This does not significantly change the results. In fact, I find that even taking this into account, the global institution's problem is approximately equivalent to the problem of minimizing $\int_{\gamma=\underline{\gamma}}^{\bar{\gamma}} y(\tau(\gamma))d\gamma$, i.e. to the problem of maximizing the reduction in aggregate demand for the fossil fuel. I document this in section 3.3.

[5]

This normative approach is in tension with the specification of the first part of the global institution's objective: the countries' aggregate private utilities. Here I am implicitly assuming that the global institution defers to the countries' revealed preferences. This is potentially problematic, since the utilities u that I take to generate the countries' revealed preferences contain a "behavioural" term, namely the tax aversion term $-\phi\tau$. If one interprets this term to capture real welfare effects, like those corresponding to the societal disruptions that would be triggered by carbon taxes, then it would be normatively appropriate to include this term in the global institution's objective, as I do here.

However, there are other empirically important effects captured by the tax aversion term, $-\phi\tau$. Firstly, it can arise from the fact that those bearing the incidence of carbon taxes tend to be well-organized and able to go on strike (Sterner et al. (2012)). Secondly, a potentially important source of carbon tax aversion is the negative bias in citizens' beliefs about the incidence from carbon taxes that they would bear, leading them to oppose carbon taxes for this reason. This theory is suggested by Douenne and Fabre (2020) who find for France: "while 70% of households are expected to benefit from this reform, only 14% think that they would". Both of these effects suggest that from a normative perspective the global institution should not defer to countries' revealed preferences. Instead, it should at least partially downweight the tax aversion term.

Similarly, the terms of trade term σy should be discarded in the normative evaluation. It is zero-sum in nature and therefore should not be counted as a welfare change. In fact, changes in the world market price of fossil fuels simply transfers money between exporters and importers.

Thus of the parameter$\gamma = \theta - \sigma - \frac{\phi}{e_d}$ we should for the purposes of normative evaluation by the global institution discard fully the term σ and at least partly discard the term $\frac{\phi}{e_d}$. Instead of venturing a guess on the difficult question of how much of the term $\frac{\phi}{e_d}$ to discard, I proceed as follows: In the main part of this paper, I use γ directly for the normative evaluation, not discarding any part of it. In appendix B I consider the other extreme case, namely where the global institution fully discards the tax aversion term $\frac{\phi}{e_d}$ in its welfare evaluations. In section 3.3 I find that the conclusions differ little between these two extreme assumptions if the global institution's annual budget is less than $100 billion.

2.1 The optimal mechanism under complete information

Under complete information, the global institution's problem is to choose a carbon tax rate $\tau(\gamma)$ for each type γ to solve the following:

Problem 1.

$$\max_{\tau(.)} \int_{\gamma=\underline{\gamma}}^{\bar{\gamma}} f(\gamma)(v(\tau(\gamma),\gamma) - \eta y(\tau(\gamma)))d\gamma$$

under the following 2 constraints:
1. exogenous budget:

$$\int_{\gamma=\underline{\gamma}}^{\bar{\gamma}} f(\gamma)t(\tau(\gamma),\gamma)d\gamma \leq M$$

where M denotes the global institution's exogenous budget.
2. voluntary participation:

$$v(\tau(\gamma),\gamma) + t(\tau(\gamma)) \geq 0$$

Lemma 3. *Under complete information, the optimal contract implements the following profile of tax rates:*
For $M \leq \eta^2 \frac{e_d}{2}$*:*

$$\tau(\gamma) = \gamma + \sqrt{\frac{2M}{e_d}}$$

The associated welfare gains w *are given by*

$$w = \eta\sqrt{2Me_d} - M$$

The resulting reduction in the use of the fossil fuel is given by:

$$\sqrt{2e_dM}$$

For $M \geq \eta^2 \frac{e_d}{2}$*:*

$$\tau(\gamma) = \gamma + \eta$$

The associated welfare gains w *are given by :*

$$w = \eta^2 \frac{e_d}{2}$$

The resulting reduction in the use of the fossil fuel is given by:

$$e_d\eta$$

Proof. See appendix A.1. □

2.2 The global institution's problem under incomplete information

Under incomplete information (or in the setting where the global institution can only condition on the tax rate τ but not the type γ), the following incentive compatibility constraint arises:

$$v(\tau(\gamma),\gamma)+t(\tau(\gamma))\geq v(\tau',\gamma)+t(\tau')\forall\gamma,\tau'$$

This is because the type γ is free to choose any γ, so its choice $\tau(\gamma)$ cannot be worse than any other τ.

Definition 2. Let us define $U(\gamma)$ as the utility gain that type γ can realize by participating relative to not participating (in which case its optimal tax rate is $\tau=\gamma$.). We thus have:

$$U(\gamma):=sup_{\tau}v(\tau,\gamma)+t(\tau)$$

By the envelope theorem, we have

$$U'(\gamma)=\frac{\partial v}{\partial\gamma}|_{(\tau(\gamma),\gamma)}$$

where $\tau(\gamma):=argmax_{\tau}\, v(\tau,\gamma)+t(\tau)$.
Using Lemma 2, we get:

$$U'(\gamma)=e_d(\tau(\gamma)-\gamma)$$

Lemma 4. *Suppose that the density function $f(\gamma)$ of the type distribution is non-decreasing. Then it is never optimal for the global institution to implement a τ with $\tau(\gamma)<\gamma$ for any γ.*

Proof. See appendix A.2. □

From the preceding Lemma 4 and $U'(\gamma)=e_d(\tau(\gamma)-\gamma)$, it follows that $U(\gamma)$ is non-decreasing. Hence there exists γ_L such that the participating types are precisely those with $\gamma\geq\gamma_L$. Moreover, the participation constraint must be binding at γ_L. Thus we obtain:

$$U(\gamma)=0\forall\gamma\leq\gamma_L$$

$$U(\gamma)=e_d\int_{\gamma=\gamma_L}^{\gamma}(\tau(\gamma)-\gamma)d\gamma\forall\gamma\geq\gamma_L$$

We also have:

$$U(\gamma)=v(\tau(\gamma),\gamma)+t(\gamma)$$

which yields:

$$t(\gamma)=-v(\tau(\gamma),\gamma)+e_d\int_{\gamma=\gamma_L}^{\gamma}(\tau(\gamma)-\gamma)d\gamma$$

Thus the global institution's problem reduces to the following:

Problem 2.

$$\max_{\tau(.)} \int_{\gamma=\underline{\gamma}}^{\bar{\gamma}} f(\gamma)(v(\tau(\gamma),\gamma) - \eta y(\tau(\gamma)))d\gamma$$

under the following constraint:

$$\int_{\gamma=\underline{\gamma}}^{\bar{\gamma}} (-v(\tau(\gamma),\gamma) + U(\gamma))d\gamma \leq M$$

where M denotes the global institution's exogenous budget and $U(\gamma)$ is given by:

$$U(\gamma) = 0 \forall \gamma \leq \gamma_L$$

$$U(\gamma) = e_d \int_{\gamma=\gamma_L}^{\gamma} (\tau(\gamma) - \gamma)d\gamma \forall \gamma \geq \gamma_L$$

Lemma 5. *Let λ denote the Lagrange multiplier associated with the budget constraint. The optimal direct mechanism satisfies for all types γ that participate:*

$$\eta - \lambda \frac{1 - F(\gamma)}{f(\gamma)} - (1 + \lambda)(\tau - \gamma) = 0$$

Proof. See appendix 2.1. □

From now on I will make the following:

Assumption 2. *γ is distributed uniformly on $[\underline{\gamma}, \bar{\gamma}]$.*

I will now define $\mu := \frac{1}{\lambda+1}$. Thus $\mu = 0$ corresponds to the case where the budget M is 0 and $\mu = 1$ to the case where the budget constraint is no longer binding[6].

Lemma 6. *For any budget $M \geq 0$, let $\mu(M) := \frac{1}{\lambda(M)+1}$, where $\lambda(M)$ denotes the global institution's shadow value of funds when its budget is M. The optimal mechanism implements the following profile of tax rates:*

$$\tau(\gamma) = \gamma + max(0, \mu(M)\eta - (1 - \mu(M))(\bar{\gamma} - \gamma))$$

Moreover, the following holds:
$\mu(0) = 0$
$\mu(M) = 1 \forall M \geq \frac{1}{2} e_d \eta(\eta + \bar{\gamma} - \underline{\gamma})$
$\mu(M)$ is strictly increasing on $[0, \frac{1}{2} e_d \eta(\eta + \bar{\gamma} - \underline{\gamma})]$
For $M \in [0, \frac{e_d \eta(\bar{\gamma}-\underline{\gamma})^2(2\eta+\bar{\gamma}-\underline{\gamma})}{6(\eta+\bar{\gamma}-\underline{\gamma})^2}]$ *the $\mu(M)$ is given as the unique* $\mu \in [0, \frac{\bar{\gamma}-\underline{\gamma}}{\bar{\gamma}-\underline{\gamma}+\eta}]$ *satisfying:*

$$\frac{e_d \eta^3 (2 - \mu)\mu^3}{6(1 - \mu)^2(\bar{\gamma} - \underline{\gamma})} = M$$

and the set of participating types is given by $[\bar{\gamma} - \eta \frac{\mu}{1-\mu}, \bar{\gamma}]$.
For $M \in [\frac{e_d \eta(\bar{\gamma}-\underline{\gamma})^2(2\eta+\bar{\gamma}-\underline{\gamma})}{6(\eta+\bar{\gamma}-\underline{\gamma})^2}, \frac{1}{2} e_d \eta(\eta + \bar{\gamma} - \underline{\gamma})]$ *the $\mu(M)$ is given as the unique* $\mu \in [0, \frac{\bar{\gamma}-\underline{\gamma}}{\bar{\gamma}-\underline{\gamma}+\eta}]$ *satisfying:*

[6] By Lemma 3, we know that the first best allocation is $\tau(\gamma) = \gamma + \eta$. With a sufficiently large budget, this allocation can also be implemented under incomplete information. For this, the global institution could simply offer a reward payment $t(\tau) := e_d \eta(\tau - \underline{\gamma})$.

$\frac{e_d}{6}\left(\mu^2\left(3\eta^2+3\eta(\bar{\gamma}-\underline{\gamma})+(\bar{\gamma}-\underline{\gamma})^2\right)-(\bar{\gamma}-\underline{\gamma})^2\right)=M$
and all types participate.
For $M>\frac{1}{2}e_d\eta(\eta+\bar{\gamma}-\underline{\gamma})$ *we have* $\mu(M)=1$ *and the first best is achieved.*

Proof. See appendix A.4. □

Consider the minimal budget required to achieve the first best: $\frac{1}{2}e_d\eta(\eta+\bar{\gamma}-\underline{\gamma})$. The width of the type space, $\bar{\gamma}-\underline{\gamma}$, captures the degree of heterogeneity between the countries. If $\bar{\gamma}-\underline{\gamma}=0$ then this formula is the same as the one we obtained for the case of complete information in Lemma 3. This of course has to be the case: $\bar{\gamma}-\underline{\gamma}=0$ means that there is no heterogeneity across the countries. More generally, the formula shows that the additional money required for achieving the first best under incomplete information exceeds the corresponding amount for complete information by the proportion $\frac{\bar{\gamma}-\underline{\gamma}}{\eta}$. This ratio foreshadows the numerical results that I will present in section 3: For coal I find that the performance of the optimal mechanism under incomplete information relative to complete information is better than that for oil. Firstly, this is because for coal the heterogeneity $\bar{\gamma}-\underline{\gamma}$ empirically turns out to be smaller. Secondly, this is because η is larger, since η is the ratio of social cost to world market price of the fuel.

The payoff heterogeneity, measured by $\bar{\gamma}-\underline{\gamma}$, also determines the range of budgets for which the optimal mechanism has only partial participation. The preceding Lemma shows that this is the case for budgets up to $\frac{e_d\eta(2\eta+\bar{\gamma}-\underline{\gamma})}{6(1+\frac{\eta}{\bar{\gamma}-\underline{\gamma}})^2}$. This is strictly increasing in $\bar{\gamma}-\underline{\gamma}$.

Why is it optimal to induce only the highest types to change their tax rates? Purely for efficiency, it would be best to have all types increase their tax rates (actually exactly by an equal amount, by Lemma 3) to reap the low hanging fruit for all types. However, this would generate large informational rents for the high types. This can be seen as follows: The slope of the reward payment functions generates incentives for countries to increase their tax rates. In order to create incentives for the low types to increase their tax rates, the global institution would therefore need to already raise the reward payments for low tax rates. But in order to also incentivise the high types to increase their tax rates, the global institution would need to pay them an amount increasing in the tax rates, in addition to the amount that it already pays the low types. It is therefore optimal for the global institution to always disproportionately concentrate its incentive power at the top of the distribution.

This is visible in the reward payment function that implements the optimal mechanism, which is characterized in the next Lemma. It shows that for low budgets, countries are rewarded in proportion to the square of the amount by which their tax rate exceeds a certain reference level τ^*. If they choose a tax rate below the reference level τ^* then they do not receive any reward payments at all. Once the budget is sufficiently large that it is optimal to induce full participation, there is an additional linear term that appears. Countries are now also rewarded in proportion to the amount by which their tax rate exceeds the reference level τ^*. As the budget M goes to the level sufficient to implement the first best, the quadratic part in the reward function converges to 0. Thus the optimal reward function converges to an affine linear function, where countries are rewarded simply in proportion to the amount by which their tax rate exceeds the reference level τ^*.

Lemma 7. *Let* $\mu(M):=\frac{1}{\lambda(M)}$*, where* $\lambda(M)$ *denotes the global institution's shadow value of funds when its budget is* M*. The optimal mechanism can be implemented by the following reward function:*

$$t(\tau)=\alpha max(0,\tau-\tau^*)+\beta(max(0,\tau-\tau^*))^2 \tag{1}$$

where α*,* β *and* τ^* *are determined from the budget* M *via the following:*
For $M\leq\frac{e_d(\bar{\gamma}-\underline{\gamma})^2(2+\frac{\bar{\gamma}-\underline{\gamma}}{\eta})}{6(1+\frac{\bar{\gamma}-\underline{\gamma}}{\eta})^2}$ *we have:*
$\alpha=0$, $\beta=\frac{1-\mu(M)}{2(2-\mu(M))}e_d$,$\tau^*=\bar{\gamma}-\frac{\eta\mu(F)}{1-\mu(F)}$

For $M \geq \frac{e_d\eta(\bar{\gamma}-\underline{\gamma})^2(2\eta+\bar{\gamma}-\underline{\gamma})}{6(\eta+\bar{\gamma}-\underline{\gamma})^2}$ *we have:*

$\alpha = \sqrt{\frac{(\eta\mu(M)-(\bar{\gamma}-\underline{\gamma})(1-\mu(M)))^2}{2-\mu(F)}}e_d$, $\beta = \frac{1-\mu(M)}{2(2-\mu(M))}e_d$, $\tau^* = \frac{\eta\mu-\sqrt{(2-\mu)((\bar{\gamma}-\underline{\gamma})(1-\mu)-\mu\eta)^2}+(1-\mu)\bar{\gamma}}{1-\mu}$

In particular, we have:

β *is strictly decreasing in* M.

α *is strictly increasing in* M.

τ^* *is first strictly decreasing in* M *for* $M \in [0, e_d(\bar{\gamma}-\underline{\gamma})^2(2+\frac{\bar{\gamma}-\underline{\gamma}}{\eta})]$ *and increasing in* M *for* $M > [e_d(\bar{\gamma}-\underline{\gamma})^2(2+\frac{\bar{\gamma}-\underline{\gamma}}{\eta}), \frac{1}{2}e_d\eta(\eta+\bar{\gamma}-\underline{\gamma})]$.

Proof. See appendix A.5 □

2.3 The limiting case where the global institution only cares about reducing emissions

In the general model studied so far, the countries' utilities (excluding climate change effects) enter the global institution's Lagrangian in two ways: Firstly, they appear in the global institution's objective function, since the global institution cares intrinsically about it. Secondly, they enter through the participation and incentive compatibility constraints. How do the results change if the first effect disappears, i.e. if the global institution's objective function only consists of the emissions reductions? Or, equivalently, what happens in the model if the social cost of carbon η goes to ∞?

It turns out that for the calibrations presented in section 3 the optimal mechanism changes very little, at least if the available budget is not much larger than what is plausible (specifically, if the annual budgets do not exceed $100 billion). Given this fact, it is interesting to study the limiting case where the global institution only cares about emissions reductions. The following Lemma gives explicit expressions for the optimal mechanism under this assumption:

Lemma 8. *Consider the limiting case where the global institution only cares about reducing emissions, which corresponds to the limit as the social cost of carbon*η *goes to* ∞. *Then we have at the optimal mechanism:*

For $M \leq \frac{e_d(\bar{\gamma}-\underline{\gamma})^2}{3}$ *the set of participating types is given by* $[\bar{\gamma} - (\frac{3(\bar{\gamma}-\underline{\gamma})M}{e_d})^{\frac{1}{3}}, \bar{\gamma}]$ *and these types choose the following tax rates:*

$$\tau(\gamma) = \gamma + (\frac{3(\bar{\gamma}-\underline{\gamma})M}{e_d})^{\frac{1}{3}} - (\bar{\gamma}-\gamma)$$

The resulting reduction in the use of the fossil fuel is:

$$\frac{e_d}{2}(\frac{3M}{e_d})^{\frac{2}{3}}(\bar{\gamma}-\underline{\gamma})^{-\frac{1}{3}}$$

Thus the ratio of the resulting reduction in the use of the fossil relative to the reduction achievable under complete information is:

$$\frac{3^{\frac{2}{3}}}{2^{\frac{3}{2}}}(\frac{M}{e_d})^{\frac{1}{6}}(\bar{\gamma}-\underline{\gamma})^{-\frac{1}{3}}$$

For $M \geq \frac{e_d(\bar{\gamma}-\underline{\gamma})^2}{3}$ *all types participate and choose the following tax rates:*

$$\tau(\gamma) = \gamma + \sqrt{2\frac{M}{e_d} + \frac{1}{3}(\bar{\gamma}-\underline{\gamma})^2} - (\bar{\gamma}-\gamma)$$

The resulting reduction in the use of the fossil fuel is:

$$e_d(\sqrt{2\frac{M}{e_d}+\frac{1}{3}(\bar{\gamma}-\underline{\gamma})^2}-\frac{(\bar{\gamma}-\underline{\gamma})^2}{2})$$

Thus the ratio of the resulting reduction in the use of the fossil fuel relative to the reduction achievable under complete information is:

$$\frac{\sqrt{2e_dM+\frac{1}{3}e_d(\bar{\gamma}-\underline{\gamma})^2}-e_d\frac{(\bar{\gamma}-\underline{\gamma})^2}{2}}{\sqrt{2e_dM}}=\sqrt{1+\frac{1}{3M}(\bar{\gamma}-\underline{\gamma})^2}-\sqrt{e_d}\frac{(\bar{\gamma}-\underline{\gamma})^2}{2\sqrt{2M}}$$

In particular, this ratio converges[7] *to 1 as M goes to ∞.*

The formula $\tau(\gamma)=\gamma+\sqrt{2\frac{M}{e_d}+\frac{1}{3}(\bar{\gamma}-\underline{\gamma})^2}-(\bar{\gamma}-\gamma)$ shows that the higher types are always induced to increase their tax rates by more. In fact, for any two participating types, their difference in tax rates is exactly twice that prevailing in the absence of the carbon pricing reward funds. This property holds here regardless of the global institution's budget.

Now let us compare this with the case studied in the preceding sections, where the global institution also cares intrinsically about the countries' utilities. There, for small budgets, this property also holds for the optimal mechanism (see Lemma 6). However, for large budgets the increases in the tax rates that the optimal mechanism induces gets more and more homogeneous across countries. Once the budget is sufficient to implement the first best, all countries are induced to increase their tax rates by the same amount, η.

This can be explained as follows: As the budget gets larger, the gains from further emissions reductions relative to the deadweight loss[8] incurred by countries get smaller. Thus the relative importance of achieving a given emissions reduction at minimal aggregate deadweight loss declines as compared to the importance of reducing informational rents.

Lemma 9. *Consider the limiting case where the global institution only cares about reducing emissions, which corresponds to the limit as the social cost of carbonη goes to ∞. The optimal mechanism can be implemented by the following reward function:*

$$t(\tau)=\alpha max(0,\tau-\tau^*)+\beta(max(0,\tau-\tau^*))^2 \tag{2}$$

where $\beta=\frac{1}{4}e_d$ and α and τ^* are determined from the budget M via the following:
For $M\leq\frac{e_d(\bar{\gamma}-\underline{\gamma})^2}{3}$ we have:
$\alpha=0$, $\tau^*=\bar{\gamma}-(\frac{3(\bar{\gamma}-\underline{\gamma})}{e_d}M)^{\frac{1}{3}}$
For $M\geq\frac{e_d(\bar{\gamma}-\underline{\gamma})^2}{3}$ we have:
$\alpha=\frac{(\frac{3(\bar{\gamma}-\underline{\gamma})}{e_d}M)^{\frac{1}{3}}-(\bar{\gamma}-\underline{\gamma})}{\sqrt{2}}e_d$, $\beta=\frac{1}{4}e_d$, $\tau^*=\bar{\gamma}-\sqrt{2}(\bar{\gamma}-\underline{\gamma})+(\frac{3(\bar{\gamma}-\underline{\gamma})}{e_d}M)^{\frac{1}{3}}(1+\sqrt{2})$

Proof. See appendix A.6. □

[7] The actual ratio for the case of oil turns out to be be less than 0.6 even for the optimistic assumption that M=$100 billion per year, as I show in section 3. In fact, the rate of convergence is ridiculously slow.

[8] I am using the term "deadweight loss" here in the spirit of the non-paternalistic interpretation of the countries' utilities implicit in the preceding sections. See appendix B for a discussion of an alternative paternalistic approach.

Interestingly, the larger the demand elasticity parameter e_d, the more strongly the optimal mechanism concentrates its incentive power on the high end of the distribution. To see why the preceding Lemma implies this, note that the reference level τ^* is strictly increasing in e_d. For budgets leading to partial participation, this means that only types above τ^* are actually induced to increase their tax rates (by Lemma 8). Since the global institution thus rewards fewer countries, it can afford to set a larger rate β for the quadratic reward payment. This explains why β is increasing in (indeed, proportional to) e_d.

3 Numerical applications

I now apply the model separately to the case of coal and oil. For all of this section I will assume that the only countries eligible for receiving the reward payments from the global institution are the non-annex 1 countries. I show results for annual budgets ranging from \$0 to \$100 billion. The rich countries had promised to provide annual funding for \$100 billion to the non-annex 1 countries from 2015 onwards for mitigation and adaptation. However, current annual funding for all global environmental institutions is only 5 billion (Stern (2023)).

3.1 Coal

3.1.1 Empirical calibration of parameters

From now on I set $p = 1$ for both coal and oil. This is simply a normalization, amounting to choosing the units of coal and oil accordingly.

$e_{d\,coal} = 0.7$, based on Keen et al. (2019) for the price elasticity of coal demand

$e_{s\,coal} = 1.3$, based on Dahl (2009) for the price elasticity of coal supply

$\eta_{coal} = 14.7$ based on a social cost of carbon of \$417 per ton of CO2 (based on Ricke et al. (2018)).[9]

$\underline{\gamma}_{coal} = 0, \bar{\gamma}_{coal} = 0.22$, based on the observation that coal subsidies are very close to 0 globally (Coady et al. 2017, Coady et al. 2019) and based on the observation that some developing countries have modest carbon prices covering coal, most notably China with \$6.3 per ton of CO2 in 2020 , which corresponds in the model to China having $\gamma = 0.22$.

$\int_{\gamma=\underline{\gamma}_{coal}}^{\bar{\gamma}_{coal}} y_{coal}(\gamma) p_{coal} = 553\, billion$ for the annual dollar amount of spending on coal by non-annex 1 countries[10]

3.1.2 Optimally rewarding non-Annex 1 countries for taxing coal combustion

From Lemma 6, we obtain that the minimal annual budget leading to full participation at the optimal mechanism is $\$5.6 billion$.

For annual budgets smaller than this, the optimal reward payment function is tangent to the x-axis at the minimal participating type, by Lemma 7, and as seen in the following plot:

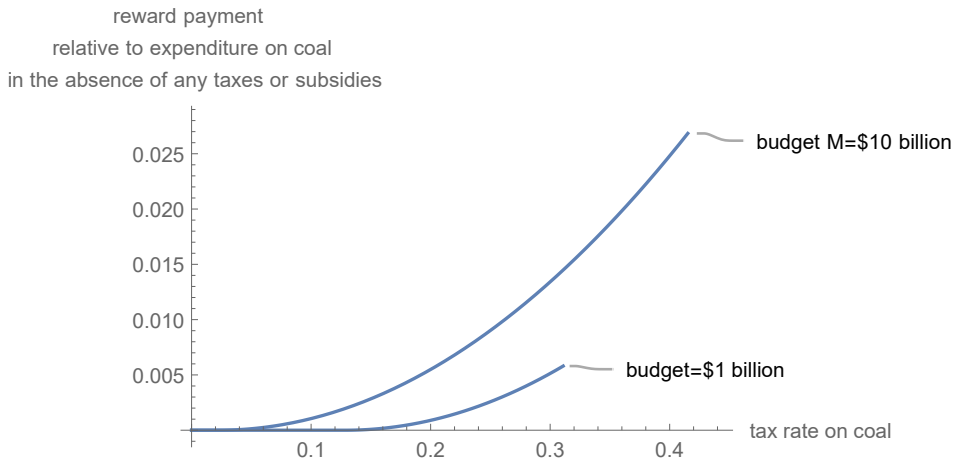
[9] This paper tries to answer a normative question: How should the global institution best be designed so as to maximize aggregate well-being? Therefore, it is appropriate to use a 0 rate of pure time preference when aggregating future climate damages.

Recall that by Lemma 1 our demand specification has the following implicit normalisation: Price at the status quo is 1. Thus η is the ratio of the social cost of a unit of coal divided by its price. To compute this we use the following: The heat content of bituminous coal is25 million Btu per short ton. Emissions from coal combustion are

79kgCO2 per GJ. Converting from GJ to Btu yields 79/0.94791 kgCO2 per million Btu. Thus emissions are 79/0.94791*25kgCO2 per short ton of coal. The average annual sale prices of bituminous coal at mines in the US in 2019 was \$58.93 per short ton. Given the assumed social cost of carbon of 417\$ per ton of CO2 leads to $\eta/p = 417 * 231.7/1000 * 25 = 14.7$. Since units are chosen such that $p = 1$, this equals η.

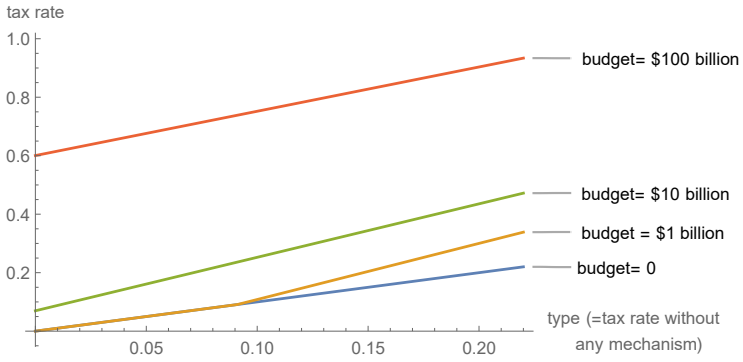
[10] 719.2\$billion is the global expected revenue from Coal Mining in 2021. Given that 77% of this is burnt in non-annex 1 countries (computed here using data from IAE (2020)), I estimate that the annual dollar amount of spending on coal by non-annex 1 countries is 0.77*719.2\$bn=553\$bn.

Figure 1: optimal reward payment functions for taxes on coal



For budgets larger than $\$5.6 billion$, there is full participation at the optimal mechanism.
The following plot shows the tax rates that end up being chosen at the optimal mechanism:

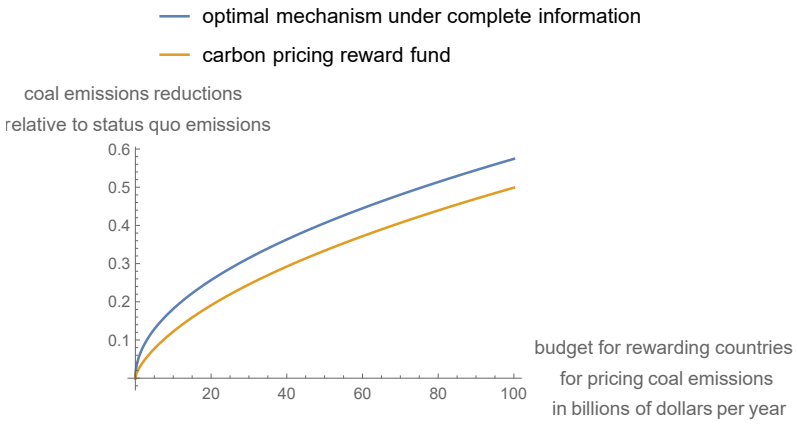
Figure 2: Tax rates on coal induced by the optimal reward function



Once there is full participation (i.e. at a budget of $\$5.6 billion$ at the global institution's disposal), further increases in the budget translate the curve of tax rates upwards. This is a property that actually holds for any distribution of types. It is merely a consequence of the fact that each country's utility is a quadratic function of its tax rate and the fact that the coeficcient of the quadratic term (in our case, half the price elasticity of demand) is the same across countries.

From now on I will refer to the optimal mechanism under incomplete information as "the carbon pricing reward fund":

Figure 3: coal emission reductions at the optimal mechanisms under complete information and incomplete information (i.e. the carbon pricing reward fund)



3.2 Oil

3.2.1 Empirical calibration of parameters

$e_{d\,oil} = 0.5$, based on Keen et al. (2019) for the price elasticity of oil demand

$e_{s\,oil} = 0.32$, based on Golombek et al. (2018) for the price elasticity of oil supply

$\eta_{oil} = 2.78$ based on a social cost of carbon of \$36 per ton of CO2 (based on Ricke et al. (2018)).[11]

$\int_{\gamma=\underline{\gamma}_{oil}}^{\bar{\gamma}_{oil}} y_{oil}(\gamma) p_{oil} = 1177\,billion$ for the annual dollar amount of spending on oil by non-annex 1 countries [12]

$\underline{\gamma}_{oil} = -1, \bar{\gamma}_{oil} = 0.1$, based on the following data from Sovacool (2017):

[11] The EPA states: The average carbon dioxide coefficient of distillate fuel oil is 429.61 kg CO2 per 42-gallon barrel (EPA 2018). The fraction oxidized to CO2 is 100 percent (IPCC 2006). The EIA states: The price of Brent crude oil, the international benchmark, averaged \$64 per barrel (b) in 2019. Assuming a social cost of carbon of \$417 per ton of CO2, we get: $\eta/p = 417 * 0.42961/64$=2.78. Since units are chosen such that $p = 1$, this equals η.

[12] World oil consumption in 2019 was 98.27M bbl/d. The eia states: The price of Brent crude oil, the international benchmark, averaged \$64 per barrel (b) in 2019. This gives that global oil consumption is \$2.296 trillion per year (in 2019). The share of that being consumed in non-annex 1 countries is 51% (computed here from data from the IAE(2020)).

Figure 4: Subsidies for Diesel in selected countries from Sovacool (2017)

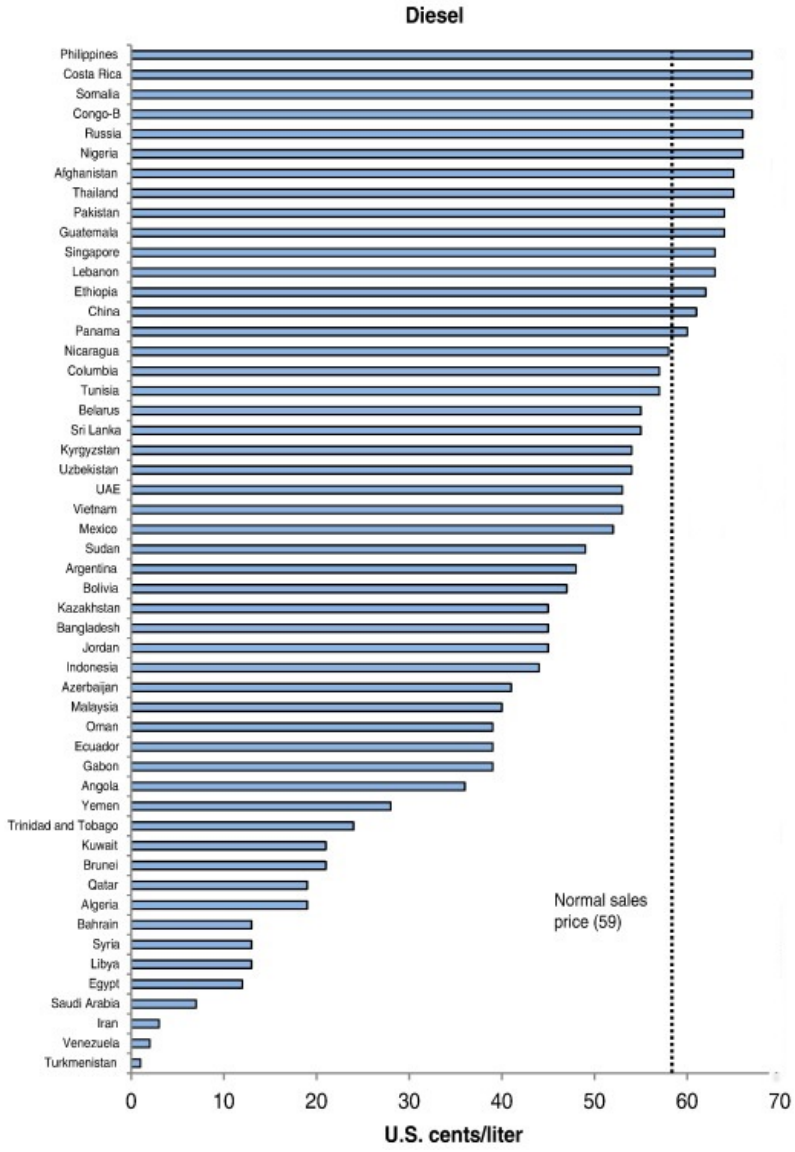
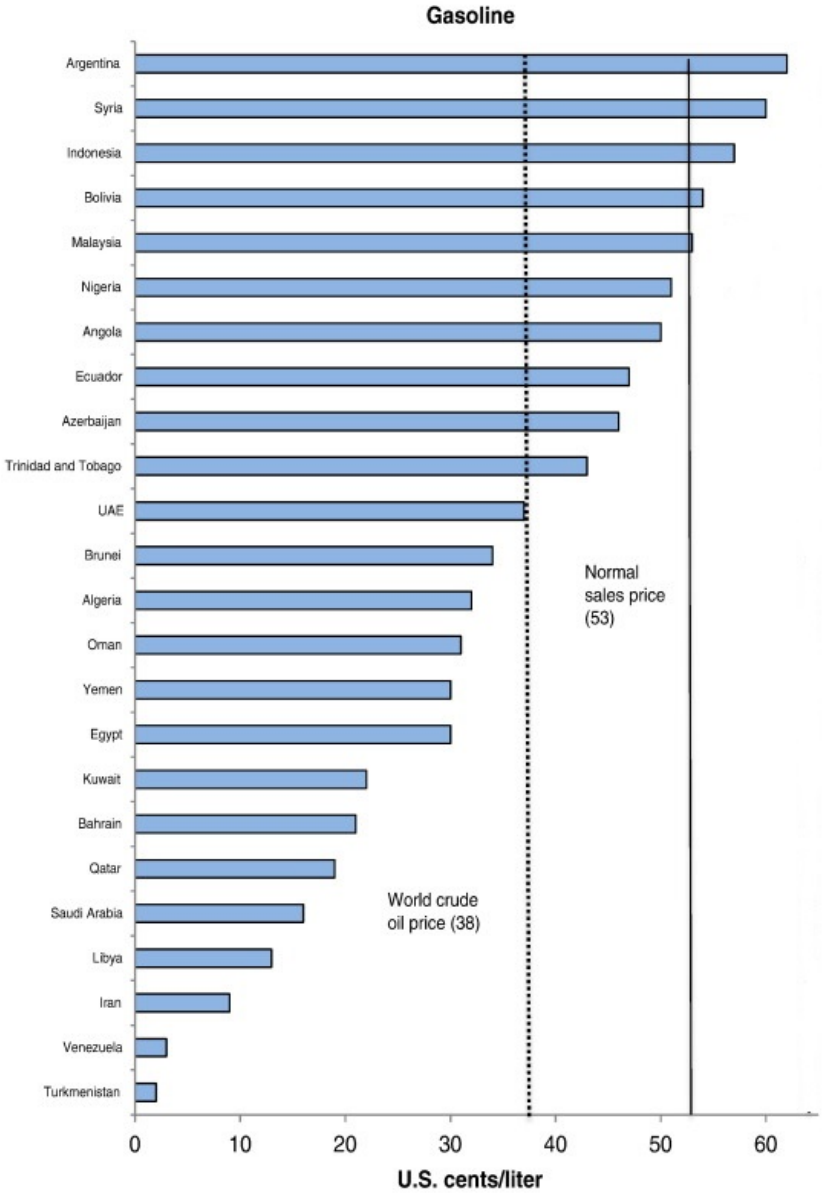


Figure 5: Subsidies for Gasoline in selected countries from Sovacool (2017)



The "normal sales price" shown here is the international spot market plus an estimate of distribution costs. The

difference between the actual sales price and the normal sales price, I take to be the fossil fuel tax rate. According to this definition, the above figures show that tax rates on oil in non-annex 1 countries range from values almost as low as -1 to values of around 0.1. This motivates the assumption that $\underline{\gamma}_{oil} = -1, \bar{\gamma}_{oil} = 0.1$.

3.2.2 Optimally rewarding non-Annex 1 countries for taxing oil combustion

From Lemma 6, we obtain that the minimal annual budget leading to full participation at the optimal mechanism is $\$58.2 billion$. This is very large because the status quo tax rates have such a wide range.

Figure 6: optimal reward payment functions for taxes on oil

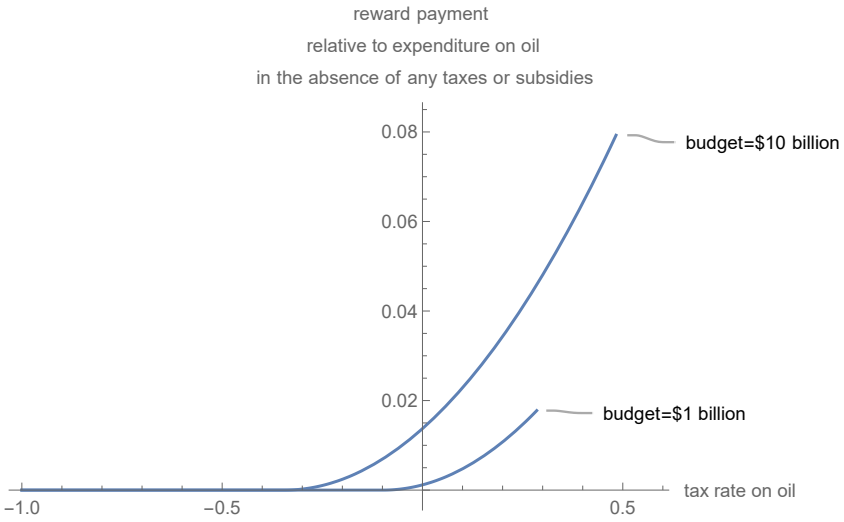


Figure 7: optimal reward payment functions for taxes on oil

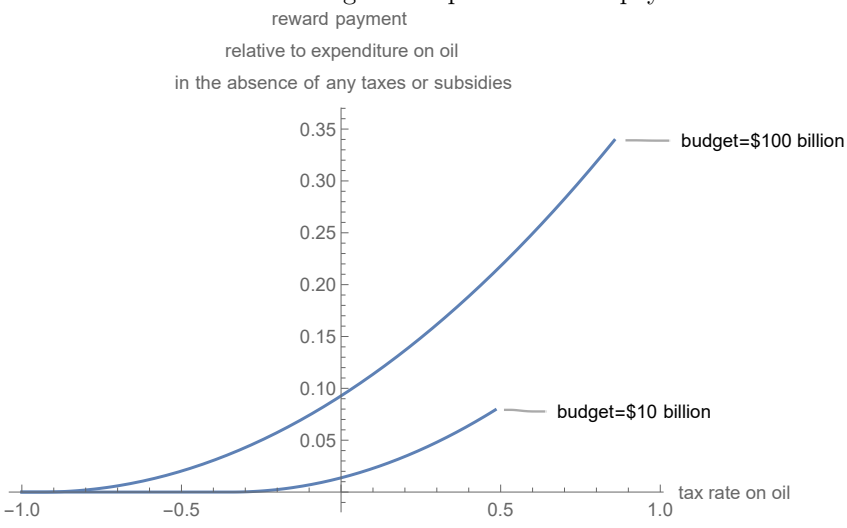


Figure 8: Tax rates on oil induced by the optimal reward function

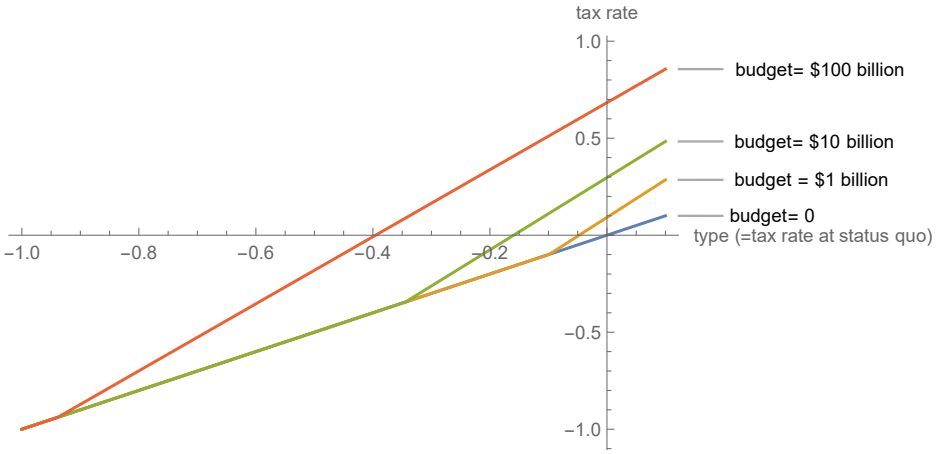
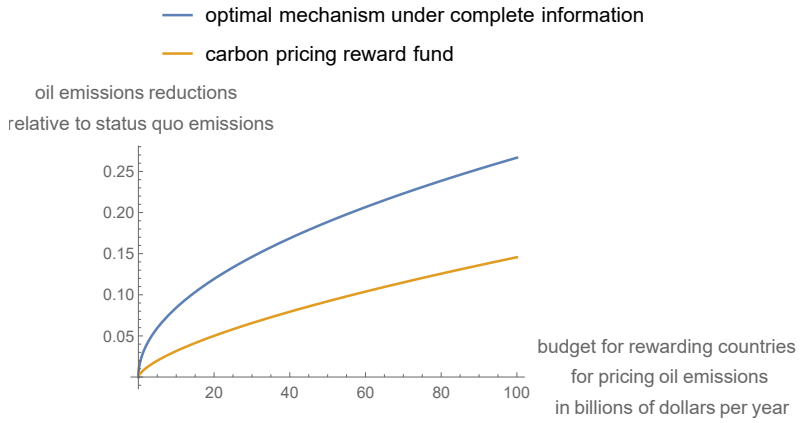


Figure 9: oil emission reductions at the optimal mechanisms under complete information and incomplete information (i.e. the carbon pricing reward fund)



3.3 Sensitivity analysis with respect to the global institution's objective

Consider a country's utility:

$$u(\tau, \theta, \phi) = B(y(\tau)) - py(\tau) - \theta y + \sigma y - \phi\tau$$

This utility is to be interpreted as corresponding to the country's revealed preference: The model assumes that the countries adjust their carbon prices so as to maximize this utility.

So far I have assumed that the global institution is non-paternalistic in the sense that it takes this utility to also be the country's welfare (not including global climate change damages). I shall now call this objective, $\int u - \eta \int y$, "global welfare under no paternalism". This underlies the results shown thus far. As discussed in footnote 5, this approach is questionable: Firstly, for the purposes of normative evaluation, the global institution should discard

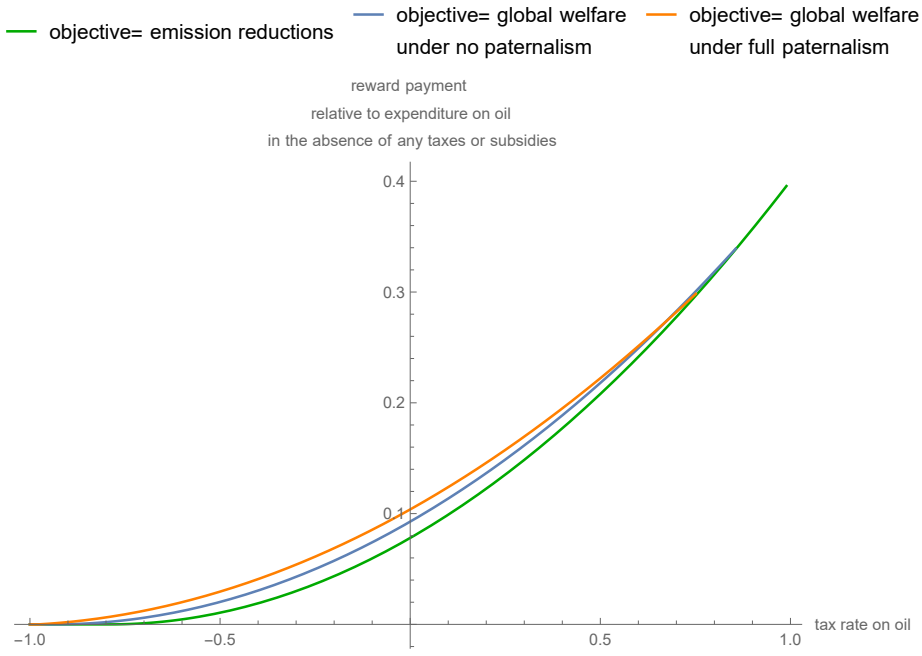
the terms of trade effect σy. Secondly, the tax aversion term $\phi\tau$ can plausibly be attributed to effects that do not reflect welfare, for example biased beliefs (Douenne and Fabre (2020)).

How much of the tax aversion term $\phi\tau$ should be discarded is difficult to know. For the results presented thus far, I have assumed that the global institution does not discard any of it. An alternative approach would be for the global institution to discard the tax aversion term completely when it evaluates the countries' welfare. I study this approach in appendix B. I call the global institution's objective in this case "global welfare under full paternalism". For tractability, I assume there that the "internalized externality", θ, is the same for all countries. Thus countries only differ in the "terms of trade effect" term σ and their tax aversion coefficient ϕ. It turns out that the global institution's problem has a solution analogous to the one in the non-paternalistic case analyzed in the main part of the paper. Qualitatively, there is a difference: The reward payment function that maximizes welfare under full paternalism is always less convex (see appendix B). This is because under full paternalism the global institution is particularly interested in increasing the tax rates of the countries that choose the lowest tax rates to reduce deadweight loss there. However, quantitatively, the difference between the optimal mechanisms under the different objective functions is relatively small, as I will show below.

For the numerical results shown in figure 10 below, I assumed that $\theta = 0.2$. This is purposefully a small value. In fact, according to IMF (2019) the average local externalities appear to be more than 0.2 in non-annex 1 countries. For this value, the optimal reward functions under full paternalism are quite close to the optimal reward functions under no paternalism, as I will explain below. It turns out that for larger values of θ this conclusion would be strengthened.

In fact, the optimal reward functions under full paternalism and under no paternalism are quite close to the optimal reward functions that the global institution would arrive at if it only cared about reducing emissions. This is shown for the case of oil with an annual budget of \$100 billion in the following graph:

Figure 10: optimal reward functions for oil for annual budget of M=\$100 billion



We see that if the global institution only cared about reducing emissions (in green) then it would concentrate its incentive power (i.e. the gradients of the reward function) even more strongly at the high end of the type distribution than if its objective is global welfare under no paternalism (in blue), as we have assumed so far.

This can be explained as follows: If the global institution only cares about reducing emissions then it considers all informational rents accruing to the countries as pure waste. This consideration pushes in favor of concentrating its incentive power at the high end of the type distribution. This consideration is only tempered by a countervailing efficiency consideration: to achieve a given amount of emission reduction at lowest cost it is best to spread it out evenly across countries. This consideration weighs in favor of equalizing the slope of the reward function over its entire domain. However, given the size of the informational rents, the first consideration is strong.

Now if the global institution cares intrinsically about the current utility of the countries that it contracts with then it does not consider informational rents accruing to the countries as being pure waste. Instead, it still values this somewhat. This dampens the force of the first consideration and thus leads to a more even spread of the incentive power. This explains why the blue curve has higher slopes than the green curve at low tax rates and lower slopes at high tax rates.

Under (full) paternalism there is a further consideration weighing in favor of creating stronger incentives for the low types to increase their tax rates: From the paternalistic perspective, the low (and even negative) tax rates chosen by the low types constitute a pre-existing inefficiency, independently of the global externalities. The lower the type, the stronger this inefficiency. The global institution wants to correct this and therefore creates incentives for the low types. This explains why the orange curve has larger slopes than the blue and green curves at low tax rates and smaller slopes at high tax rates.

Figure 10 suggests that the optimal reward functions depend relatively little on which of the three objectives the global institution pursues. In fact, this dependence on the objective is much weaker still for smaller annual budgets. Intuitively, this is because for lower budgets the marginal value of relaxing the global institution's budget constraint rises and this means that reducing informational rents becomes an even more important consideration under each of the objectives. For coal the optimal reward functions depend even less on which of the three objectives the global institution pursues.

Normatively, the appropriate objective arguably lies somewhere between "global welfare under full paternalism" and "global welfare under no paternalism". However, the results just explained suggest that the objective "emission reductions" generates a decent approximation. Taking the objective to be "emission reductions" has the advantage of making the optimal mechanism much easier to understand, as shown by Lemma 8 and Lemma 9 that give explicit formulae for the coefficients of the optimal reward payment function. For this reason I propose to use the objective "emission reductions" for designing carbon pricing reward funds in practice, as I will detail in the next section.

4 Implementing the mechanism in practice

I now propose an operational method for implementing carbon pricing reward funds in practice, using the solution obtained in section 2.3 to define a reward payment function. The method is based on five design choices that I motivate in the sub-sections afterwards.

4.0.1 A proposal

For each specific type of fossil fuel (coal for electricity, coal for steel production, oil for transportation, oil for heating, etc.) a separate carbon pricing reward fund is established. Donors can give money to each of these funds separately. Each fund uses its money each year according to the procedure that I now define.

Before the beginning of each year, the optimal mechanism is computed in the model presented in section 2.3, using best available estimates of price elasticities of demand for the fossil fuel, assumed to be constant across

countries. Specifically, this means that the fund announces before the beginning of the year that it will at the end of the year reward countries as a function of their tax rate τ on the fossil fuel[13] according to the following formula:

$$t(\tau) = \alpha max(0, \tau - \tau^*) + \beta(max(0, \tau - \tau^*))^2 \tag{3}$$

For each non-annex 1 country i, the fund computes an estimate of $y_i(0)$, the amount of the fossil fuel that the country i would use if it set $\tau = 0$ and if moreover, it did not implement any other policies affecting the use of the fossil fuel. The fund announces that country each country i will receive the following reward payment at the end of the year, depending only on its choice τ_i of the tax rate on the fossil fuel:

$$y_i(0)t(\tau_i) \tag{4}$$

The fund commits to sticking to the value of τ^* that it computed before the beginning of the year. For the other two parameters, α and β, it announces that it will scale them by an equal factor such that at it ends up disbursing exactly its budget available for the year[14].

4.0.2 Motivations for the proposal

A now go through a list of design choices that are implicit in the above proposal. I justify each of these choices.

Using the limiting case where the global institution only cares about reducing emissions In section 3.3, I reported on results showing that if the global institution only cares about reducing emissions then the optimal mechanism will be hardly different from the case where the global institution weighs emissions reduction against current day welfare using the social cost of carbon estimates explained in section 3.1.1. The former assumption has the advantage of making the optimal mechanism much easier to understand, as shown by Lemma 8 and Lemma 9 that give explicit formulae for the coefficients of the optimal reward payment function. I therefore propose to use this assumption.

Creating a separate carbon pricing reward fund for each type of fossil fuel In the numerical applications of section 3 I considered the case where there is a separate carbon pricing reward fund for coal and oil. In practice, I propose to further differentiate and to instead create a separate carbon pricing reward fund for each sub-type of fuel (oil used for heating, oil used for transportation, coal used for electricity generation, coal used for steel production etc.). This architecture is likely to lead to easier implementation and possibly also to greater welfare gains, as I will now briefly argue.

In practice, countries differentiate their fossil fuel taxes and subsidies according to the sub-type of fuel. Under the architecture with separated carbon pricing reward funds proposed here, the model can be directly applied. If instead a single carbon pricing reward fund was used for all types of coal or oil, then application of the model would require some way of aggregating the different subsidies/taxes across the different sub-types.

Such an approach would amount to contracting jointly on the different tax rates. The following question arises: Is there any mechanism contracting jointly on the different tax rates that achieves strictly greater welfare than that obtained by optimally contracting on each of the variables separately and optimally splitting the budget between these separate mechanisms? I conjecture that the answer to this question is no as long as the correlation between the different type dimensions is sufficiently strong. In a simplified model with discrete 2 by 2 type spaces this claim

[13] More, precisely, this could be the average tax rate that the countries has in place over the year.

[14] Equivalently, we have: the ratio $\frac{\alpha}{\beta}$ would be fixed according to values generated by the model. (By Lemma 9, we have $\frac{\alpha}{\beta} = 0$ for $M \leq \frac{e_d(\bar{\gamma}-\underline{\gamma})^2}{3}$ and $\frac{\alpha}{\beta} = 2\sqrt{2}((\frac{3(\bar{\gamma}-\underline{\gamma})}{e_d}M)^{\frac{1}{3}} - (\bar{\gamma}-\underline{\gamma}))$ for $M \geq \frac{e_d(\bar{\gamma}-\underline{\gamma})^2}{3}$). The absolute values of α and β would then be set at the end of the year (when the reward payments for that year are actually made) so as to exactly achieve budget balance.

can be proven in full generality based on the results from Armstrong & Rochet (1999)(where "sufficiently strong correlation" means that their equation (7) holds). In the context of the current paper, the condition "the correlation between the different type dimensions is sufficiently strong" means roughly: "countries having a relatively high tax rate on one fossil fuel typically also have a relatively high tax rate on the other fossil fuels".

Now suppose that this empirical condition is satisfied and also suppose that the conjecture holds. Then we could conclude that nothing can be gained by jointly contracting on the different tax rates. Now this would imply that rewarding countries based on an aggregate index of their fossil fuel tax rates could not lead to strictly higher welfare. I will now argue that it might actually lead to strictly lower global welfare.

Lemma 9 shows that the optimal reward payment function depends strongly on the elasticity of demand for the sub-type of fuel. These elasticities vary substantially across sub-types of fuel[15]. Lemma 9 shows that the higher the elasticity parameter, the more strongly the global institution should concentrate its incentive power on the high end of the type space, as I explained in the paragraph after the Lemma. This suggests that lumping several sub-types of fuel together in a single carbon pricing reward fund might lead to significant welfare losses.

Allowing donors to earmark monetary contributions to each of the carbon pricing reward funds I have proposed here a separated architecture for the carbon pricing reward funds, in the sense that donors can earmark their donations to the different carbon pricing reward funds. Alternatively, one could consider a unified architecture, where donors could only donate to a central global institution, which would then distribute its available budget between the different carbon pricing reward funds according to some rule. For example, it could use the model to compute the unique way to split a given budget between the different carbon pricing reward funds so as to maximize global welfare.

An advantage of the separated architecture is that it enhances donors' incentives to contribute funding. Specifically, oil importers benefit from lower world oil prices. They therefore benefit particularly from donations to the oil carbon pricing reward funds, since these lower the world market price of oil. In Stern (2021) I find in a related model that this incentive effect always dominates the efficiency losses that come from sub-optimal splitting of funding across different funds.

Estimating the emissions in the hypothetical situation without any taxes or subsidies I have proposed here to make the reward payment to a country proportional to its counterfactual use of the fossil fuel in the absence of any government intervention. This ensures ensures that countries' with equal types (i.e. equal values of γ) face equal incentives, irrespective of their "size", where "size" here means their counterfactual use of the fossil fuel in the absence of government intervention. Formally, this equalization of incentives for countries with equal types but differing sizes is necessary for optimality if sizes and types are independently distributed.

It seems feasible to define a standardized methodology for estimating each country's fossil fuel use in the absence of any government intervention. It should take into account not only the explicit carbon prices but also other policies that reduce emissions. The goal for the methodology should be to generate as accurate as estimate as possible of the expenditure on the fossil fuel that would arise in the absence of any government intervention. This would be important to ensure that the mechanism does not in any way undermine countries' incentives to pursue other policies to reduce emissions.

Ensuring budget balance I have proposed here that the carbon fund would commit to stick to the value of the reference level τ^* that it computes before the beginning of the year based on the model but that it would adjust the other two parameters, α and β, so that it ends up exactly using up its budget for the year when it makes the

[15]For example, Douenne (2018) finds for the case of France that "the median household reacts significantly to transport fuel prices with an uncompensated price elasticity around -0.45, and to a lesser extent to housing energy prices with an elasticity of -0.2."

reward payments at the end of the year. As a result, when a country chooses its carbon tax at the beginning of the year, it would certainty about whether it will get any reward payment or not. How much reward payments it will receive will depend on how large β ends up being, which depends on the other countries' choices.

The approach I have defined here has the advantage of avoiding cases of disappointment where a country hopes to get reward payments but ends up not getting any. This is ensured by having the global institution commit to the value of τ^* that it uses for a given year before that year begins. Of course this means that all the uncertainty about the actual reward payments resides in the rates α and β at which countries are rewarded. However, it seems better to concentrate the uncertainty at the intensive rather than the extensive margin.

5 Comparing Carbon Pricing Reward Funds with project-based contracting

Current global environmental institutions predominantly use project-based contracting. I now present a model making favorable assumptions for this approach[16].

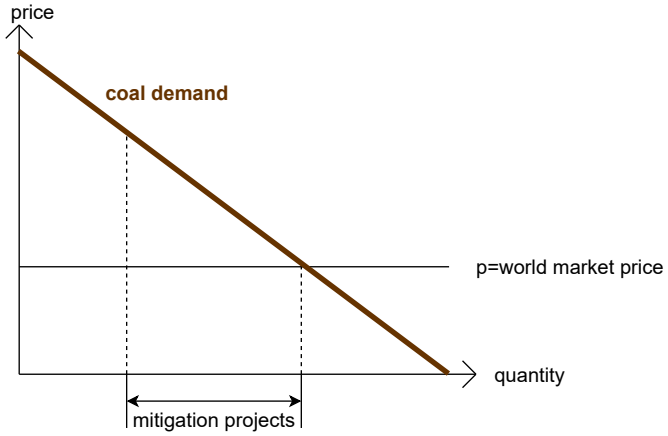
5.1 An optimistic model of project-based contracting

Consider the following model, which can be viewed as distilling the functioning of an idealized version of the Clean Development Mechanism:

All private actors can submit "projects" for emissions reductions. The global institution perfectly identifies the projects that would not be implemented without its support. To these projects, the global institution offers a uniform payment per certified emissions reduction.

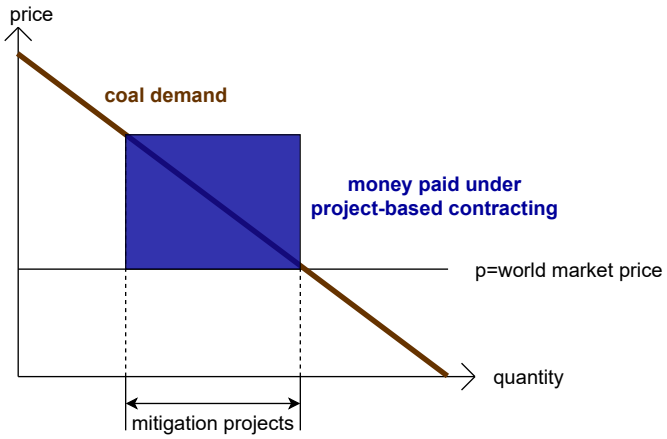
All changes in behaviour that an increase in the tax on the fossil fuel would induce can be certified as projects. This assumptions is favorable for the project-based approach. In reality, only some projects such as substitution to renewable energy and energy efficiency improvements can be certified, whilst individual behaviour change cannot.

Given that our specification leads to a linear demand function for fossil fuel by Lemma 1, we can represent the project-based contracting as follows. The mitigation projects funded are the marginal ones:

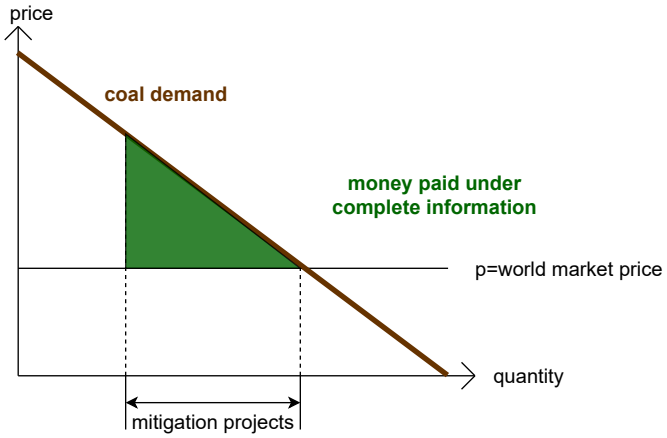


They all receive the same amount of compensation per emission reduction:

[16]Importantly, I assume that under the project-based contracting governments do not adjust their tax policies. Stern (2023) studies the endogenous government response to project-based contracting. The results would, when applied to the model of the current paper imply that project-based contracting would have no effect at all once governments anticipate it and adjust their tax policies accordingly.



The minimal required compensation that would be required under contracting with complete information is the the green area:



From this we deduce:

Lemma 10. *For any given budget, the ratio of emissions reductions achieved under project-based contracting over the emissions reductions that could be achieved under complete information is* $\sqrt{\frac{1}{2}}$.

Proof. Suppose that the available budget for project based contracting is M. By Lemma 3, the reduction in the combustion of the fossil fuel is $\sqrt{2e_d M}$. Now switching from contracting under complete information to project-based contracting is analogous to dividing the available budget by 2, given that now half of money gets paidd out as rents, as illustrated in the above diagrams. Given that $\sqrt{2e_d M}$ is the emission reduction resulting under contracting with complete information, the emission reduction resulting under project-based contracting is $\sqrt{2e_d \frac{M}{2}} = \sqrt{\frac{1}{2}}\sqrt{2e_d M}$. □

5.2 At what level of funding would the creation of a carbon pricing reward fund be valuable?

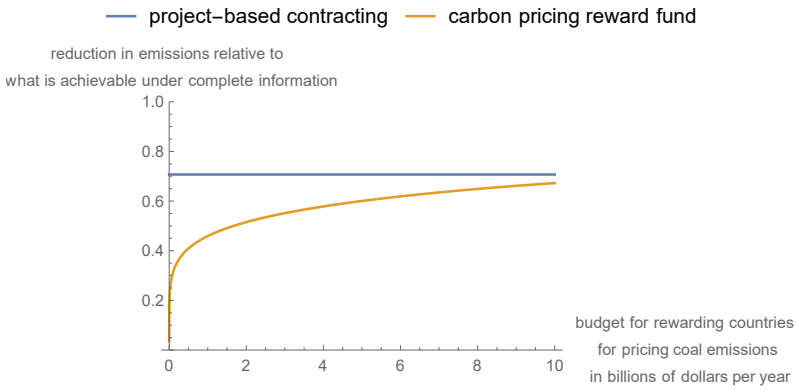
So far, global environmental institutions such as the Clean Development Mechanism (CDM) and the Green Climate Fund have relied on project based approaches to climate change mitigation. An advantage of project based ap-

proaches is that they can reduce informational rents. For example, the CDM only certifies projects that it assesses would not have been implemented in the absence of the financing that it provides. It pays projects proportionally to the amount of emissions reductions that they achieve. The model presented in section 5.1 is meant to capture exactly this. In our model, the marginal abatement cost curve is linear. This implies that the CDM ends up paying half its budget as informational rents and using its other half as if it was contracting optimally under complete information. As a result, Lemma 10 showed that the ratio of emissions reductions achieved relative to what would be achievable under complete information is $\sqrt{\frac{1}{2}} \approx 0.7$. This holds regardless of the level of the budget.

However, this number is based on the assumption that all the CDM projects are actually additionally, an assumption that has been called into question by empirical studies such as Dechezleprêtre et. al (2014). Thus in reality, the ratio of emissions reductions achieved relative to what would be achieved under complete information is for the CDM likely to be much lower than $\sqrt{\frac{1}{2}} \approx 0.7$.

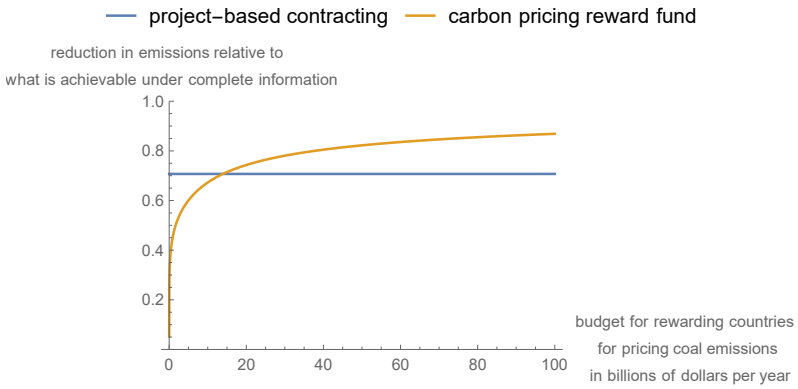
Even under this optimistic assumption about project-based contracting, we see that the carbon pricing reward fund comes close to it in the case of coal, even for relatively small budgets:

Figure 11:



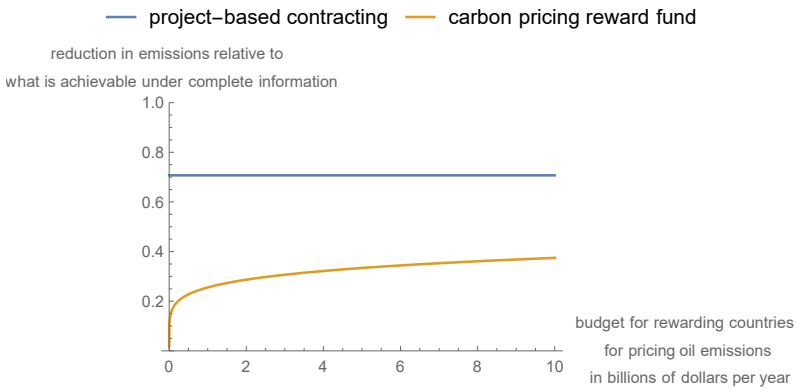
Zooming out:

Figure 12:



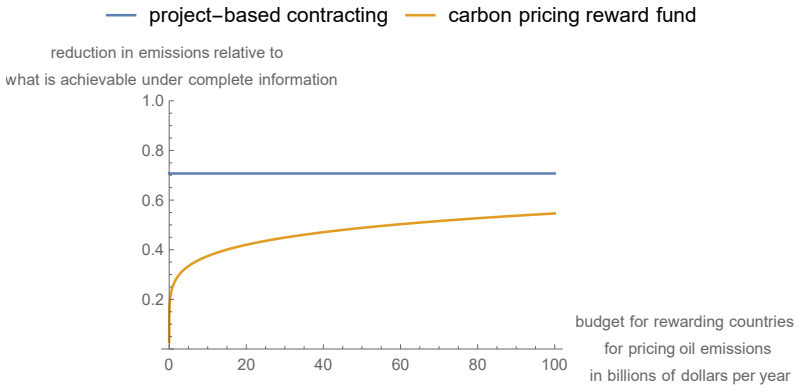
For oil we find:

Figure 13:



Zooming out:

Figure 14:



If the world were to mobilize additional funding for global climate change institutions in the order of the existing annual amounts (5 billion), then it might be optimal to direct this additional funding to newly established carbon pricing reward funds.

There are further considerations supporting this conclusion: For the carbon pricing reward funds, the main administrative costs are roughly fixed: The countries' carbon prices need to be measured and the the counterfactual fossil fuel use in the scenario without any carbon taxes need to be estimated (see section 4.0.2). For project based approaches like the Clean Development Mechanism and the Green Climate Fund, the administrative costs are roughly proportional to the number of projects, as each project needs to be assessed.

5.3 What are the optimal roles of project-based contracting and policy-based contracting (specifically, carbon pricing reward funds)?

The conclusion we have arrived at thus far is that in sectors where mitigation is determined by the price of fossil fuels, carbon pricing reward funds will likely outperform project-based contracting as soon as the world mobilizes substantial amounts of additional funding. This conclusion holds for coal even under the very optimistic model of project-based contracting presented in section 5.1. For oil it is likely to hold once one takes into account the difficulties for project-based contracting not taken into account in the model, as I argued in section 5.2.

However, there are some economic sectors where mitigation is not determined by carbon prices or other simple policy parameters such as subsidy rates for renewables. For example, governments make direct decisions about investments and pricing of public transportation and this will affect emissions. Global environmental institutions such as the Green Climate Fund currently use a part of their budget for paying governments for expanding public transportation. This is a case where project-based contracting will continue to be optimal, even in the presence of well-funded carbon pricing reward funds.

Thus we are led to the following tentative policy conclusion: If and when the world mobilizes substantial additional funding for creating positive incentives for developing countries to reduce emissions, it will be optimal to rely entirely on carbon pricing reward funds in sectors where mitigation is primarily determined by fossil fuel prices. It is likely best to create a separate carbon pricing reward fund for each specific type of fossil fuel (coal for electricity, coal for steel production, oil for transportation, oil for heating...). The existing global institutions would then continue to engage in project based contracting, but focus entirely on the sectors (e.g. expansion of public transportation) where there are policy dimensions other than carbon pricing that are of importance.

6 Towards a dynamic analysis

The formal analysis so far has been static. We have only considered mechanisms that condition each country's reward payments in a given year only on the country's tax rates in that year. A natural way to relax this restriction is to allow the global institution to also condition the reward payment to each country on its tax rates before the start of the mechanism. These prior tax rates contain valuable information about countries' types and would allow the global institution to reduce countries' informational rents. The ex post optimal mechanism will reward countries less if they had high fossil fuel taxes in the past, since this allows a reduction in the rents accruing to the "high types", i.e. the countries with a preference for high fossil fuel taxes.

Would this be problematic? Countries with high past fossil fuel taxes might regret this. The mechanism might be viewed as "unfair" since virtuous past behaviour leads to lower reward payments in the present. This might undermine the legitimacy of the global institution and also of the mechanisms that might be used to fund it. Setting aside such fairness/legitimacy considerations, the following question arises: Purely in terms of incentives, is the feature just described problematic?

If one views the design of each such global institution in isolation, one might be led to conclude that no perverse incentives arise: By construction, the past decisions are immutable when the new instutions are created.

However, a full assessment needs to take into account the larger context that we are in: If global institutions of the kind proposed in this paper have a chance of getting created for rewarding carbon pricing as considered here, then there might also be a chance that similar such institutions get created for rewarding other policies with positive global externalities. For example, it has been proposed that a new global institution be created that would reward countries based on their score in the Global Health Security Index which evaluates countries' policies and regulations in terms of how well they contribute to preventing future global pandemics (CSIS et al. (2020)).

The world is therefore faced with the following question: Should new global reward payment funds in general condition their reward payments on countries' past actions or not? It is instructive to compare the welfare implica-

tions of two extreme approaches: 1) *never* condition on countries' past actions vs 2) *always* condition on countries' past actions. Of course, the world could instead settle for some intermediate approach between these two extremes. However, a welfare comparison of the extremes can help inform a case-by-case assessment of the question as to whether reward payments should be conditioned on past actions or not. In fact, whenever a new global institution conditions reward payment on countries' past actions this gives countries rational reasons to believe that this approach will also be more likely to be adopted for future global institutions.

Motivated by these observations, this section develops a model that aims to quantify for a given global institution the welfare implications of being created in a scenario where the world has settled on approach 1) of *never* conditioning on countries' past actions vs approach 2) *of always* conditioning on countries' past actions.

6.1 A dynamic model

Time runs discretely (we will take the continuous time limit a bit further down). At the beginning of each period, there is a probability λ that a global institution is established, conditional on it not having been established before. If the global institution does not exist then only the countries move in the period. They all simultaneously decide on their tax rates for that period. If the global institution exists then after each period there is a probability κ that the global institution is dissolved and then the game ends (or starts anew, which formally comes to the same.).

If the global institution is created, then from that period onwards the game unfolds as follows. In each period, the global institution has a fixed budget M to be fully used during that period. At the beginning of each period, the global institution announces how it will split up its budget M as a function on the profile of the countries' tax rates. Formally, the global institution chooses a profile $(t_i((x_j)_{j\in I})_{j\in I}$ of functions under the constraint that $\int_{i\in I} t_i((x_j)_{j\in I}) = M$.

Then the countries all move simultaneously. They decide on their tax rates for the period. Each country has an "ex ante type", denoted as in the static model by γ. Before the creation of the global institution, a country's flow utility (relative to not participating and thus choosing $\tau = \gamma$) is like in the static model:

$$-\frac{1}{2}e_d(\tau-\gamma)^2 + t$$

Recall that we obtained this in Lemma 2 on the basis of the micro-foundations given at the beginning of section 2.

Once the global institution is created, the countries' types adjust. A country's "ex post type" is equal to $\tilde{\gamma} = \gamma + \gamma^{\#}$, where $\gamma^{\#}$ is a random variable. For now, I will not need to specify how $\gamma^{\#}$ is distributed but in the next section 6.1.2 I will assume that $\gamma^{\#}$ is independent of γ.

A country of ex post type $\tilde{\gamma}$ choosing a tax rate τ and receiving a transfer t receives the following flow utility:

$$-\frac{1}{2}e_d(\tau-\tilde{\gamma})^2 + t$$

The motivation for this is as follows: The creation of the global institution prompts each country to re-evaluate its subjective benefits and costs associated with different tax rates on the fossil fuel. For example, it leads to a public discussion of carbon taxes the outcomes of which are uncertain. This motivates the assumption that the preference shock $\gamma^{\#}$ is unknown to the country itself before the creation of the global institution.

A country's objective is the discounted integral of the flow utility. All countries use the same fixed rate r to discount future utility. We are hereby and for the rest of the paper switching to the continuous limit in the time dimension. Above, I introduced the discrete time version of the model to make clear that in each period the global institution moves first and then all the countries move simultaneously. Now that we are switching to the continuous time limit, we will need to remember that "at each instant the global institution moves first".

The global institution's objective is to maximise emission reductions. This assumption greatly simplifies the analysis. It is justified by the findings from section 3.3. Thus from now on we will mean by "welfare" simply "aggregate emission reductions".

The global institution discounts welfare with its own discount rate δ which is allowed to differ from the interest rate r. The intended interpretation is as follows. The global institution uses a pure time discount rate of 0. The only reason for weighting earlier emissions more heavily is that hastening emissions also hastens global warming and thereby increases overall damages.

6.1.1 Perverse incentives if reward payments are conditioned on the change in tax rates relative to their values just before the creation of the global institution

A natural kind of mechanism that might be considered if reward payments are conditioned on past tax rates is to reward countries entirely as a function of the increase in their tax rates relative to their tax rates before the creation of the global institution. I will later provide sufficient conditions for there to be a time-consistent such mechanism (see Lemma 12). Time-consistency here means that if countries expect the global institution to use such a mechanism then it will be optimal ex post for the global institution to indeed use it. The following Lemma quantifies the perverse incentive effects that would arise in the time before the global institution is created:

Lemma 11. *Suppose the global institution rewards countries on the basis of their change in tax rates relative to the point in time just before its creation. Suppose the mechanism is such that all countries end up participating. Then we have: Before the global institution's creation, countries set their tax rates below what they would do in the absence of any mechanism. During the global institution's existence, they set them above what they would do in the absence of any mechanism. These two effects compare as follows:*

$$\frac{\textit{tax } \textbf{\textit{decrease}} \textit{ caused by the global institution } \textbf{\textit{before}} \textit{ its existence}}{\textit{tax } \textbf{\textit{increase}} \textit{ caused by the global institution } \textbf{\textit{during}} \textit{ its existence}} = \frac{\lambda}{r+\kappa}$$

$$\frac{\textit{annual emission } \textbf{\textit{increase}} \textit{ caused by the global institution } \textbf{\textit{before}} \textit{ its existence}}{\textit{annual emission } \textbf{\textit{decrease}} \textit{ caused by the global institution } \textbf{\textit{during}} \textit{ its existence}} = \frac{\lambda}{r+\kappa}$$

$$\frac{\textit{expected discounted emission } \textbf{\textit{increase}} \textit{ caused by the global institution } \textbf{\textit{before}} \textit{ its existence}}{\textit{expected discounted emission } \textbf{\textit{decrease}} \textit{ caused by the global institution } \textbf{\textit{during}} \textit{ its existence}} = \frac{\delta+\kappa}{r+\kappa}$$

with the following meaning of the parameters:

parameter	*meaning*
λ	*yearly probability that the global institution is created*
r	*market interest rate, with which the countries discount future payoffs*
κ	*yearly probability that the global institution is dissolved*
δ	*discount rate used by the global institution to aggregate emission reductions into time-neutral welfare*

Proof. See appendix C.1. □

Consider the results from Lemma 11: During its existence, the global institution causes emisssion reductions. Before its coming into existence, perverse incentives arise that create emission increases. Once we aggregate these emissions over time using the discount rate δ, the ratio of the second effect relative to the first is given by the following formula:

$$\frac{\delta+\kappa}{r+\kappa}$$

Interestingly, this formula does not depend on λ, the annual probability of the global institution being created. This can be explained as follows. λ has two effects. On the one hand, a high λ means that for each country the expected gain from distorting downwards its tax rate is large: That year's tax rate will likely be the basis for the reward payment function that it will face during the entire time that the global institution will exist. Given the quadratic cost of distorting its tax rate, the country's optimal distortion of its tax rate is proportional to λ. On the other hand, a high λ means that this effect has typically little time to play out. In fact, the expected time until the global institution is created is $\frac{1}{\lambda}$. As a result, these two effects exactly cancel each other out.

6.1.2 Sufficient conditions for the ex post optimal mechanism to condition reward payments on the change in tax rates relative to their values just before the creation of the global institution

Definition 3. Consider the static model with type distribution according to the ex post part of the type (i.e. the random variable $\gamma^{\#}$). Consider the optimal mechanism under incomplete information. Let $R(M)$ denote the corresponding aggregate amount of rents and $X(M)$ the corresponding aggregate emission reductions.

Assumption 3 (The Large Enough Rents Assumption)**.** $\frac{R(M)}{M-R(M)} > \frac{\lambda}{r+\kappa}$.

The case where the the Large Enough Rents Assumption does not hold leads to puzzling technical difficulties that place it outside of the scope of the current paper.[17] I will illustrate in section 6.2.2 that at least in the case of the global institution rewarding countries for taxing oil combustion the assumption is plausibly valid.

Lemma 12. *Suppose the Large Rents Assumption 3 holds and suppose that the distribution of $\gamma^{\#}$ conditional on γ does not depend on γ. Then the following mechanism is time consistent:*

The global institution uses a reward payment function of the form $t(\tau_2 - \tau_1)$, where τ_2 is the country's current tax rate and τ_1 is the country's tax rate just before the global institution is created. The function t corresponds to the optimal mechanism (under incomplete information) in the static model model with a type distribution given by the random variable $\gamma^{\#} + \frac{\lambda}{r+\kappa}\frac{X(M)}{e_d}$.

Given this reward payment function, the country of ex ante type γ has a unique best action before the creation of the global institution, namely to choose the tax rate $\gamma - \frac{\lambda}{r+\kappa}\frac{X(M)}{e_d}$.

Proof. See appendix C.2 □

6.2 Illustrative numerical applications

6.2.1 The "perverse incentive effect"

By Lemma 11, the perverse incentive effect accruing before the global institution's creation cancels the proportion $\frac{\delta+\kappa}{r+\kappa}$ of the welfare gains that the global institution causes during its existence. Suppose that $\delta = 0$, i.e. suppose that what matters for global welfare is only the eventual aggregate amount of emissions. Suppose that the typical lifetime of a global institution of this kind equals that of the duration of the Kyoto Protocol's 2 commitment periods (2008-2012,2012-2020), i.e. 12 years, suggesting $\kappa = \frac{1}{12} = 0.08\bar{3}$. For an interest rate of $r = 0.05$ we get:

$$\frac{\delta+\kappa}{r+\kappa} = 0.625$$

Thus by Lemma 11, the perverse incentive effect cancels 62.5% of the welfare gains.

[17] In this case there are no pure strategy equilibria and finding the mixed strategy equilibria seems very difficult. It seems plausible that at the mixed strategy equilibria the overall efficiency losses accruing before the coming into existence of the global institution will be larger than in the case of pure strategy equilibria. In fact, the perverse incentives will still be present and in addition to that the randomization of strategies creates additional inefficiencies.

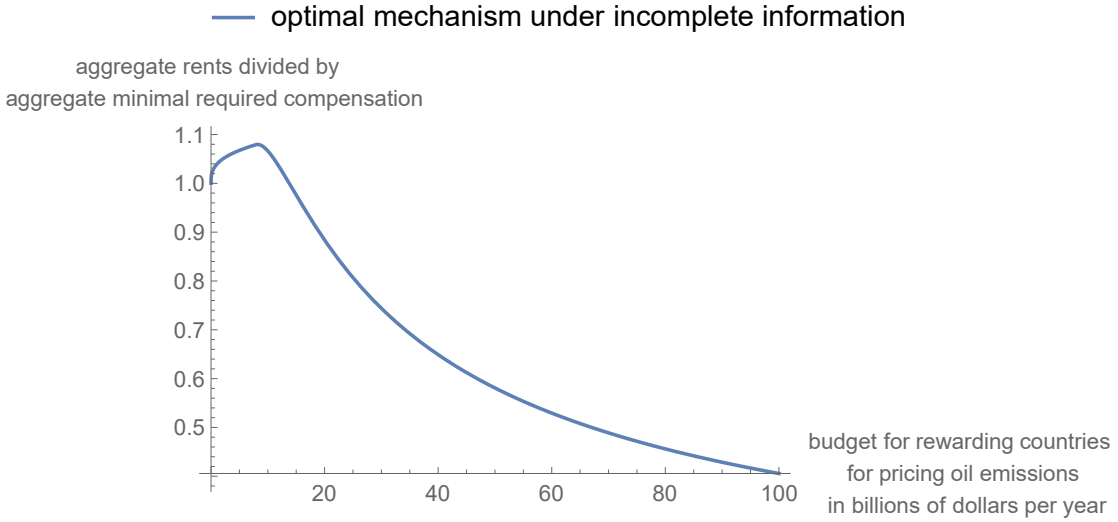
6.2.2 The "Large Enough Rents Assumption"

I will now argue that the Large Enough Rents Assumption plausibly holds for the case of a global institution rewarding countries for taxing oil combustion. Recall that this assumption is the following condition:

$$\frac{R(M)}{M - R(M)} > \frac{\lambda}{r + \kappa}$$

Consider first the left hand side. Here M denotes the global institution's budget, $R(M)$ denotes the aggregate rents accruing to countries at the optimal mechanism in the static model with types distributed according to the component $\gamma^{\#}$ of the countries' ex post types. Recall that a country ex post type equals $\gamma + \gamma^{\#}$, where γ is its ex ante type which equals by Lemma 2 its (ad valorem) tax rate in the absence of any mechanism.

Consider the case of oil. We assumed γ to be distributed uniformly on $[0, 1.1]$, as explained in section3.1.1. Now suppose that $\gamma^{\#}$ is also uniformly distributed but on an interval of 20% the width of $[0, 1.1]$. Under this assumption, the following curve plots the expression $\frac{R(M)}{M-R(M)}$ as a function of M:



Suppose $r = 0.05$, $\kappa = 0.08\bar{3}$ and $M \leq \$20 billion$. Then the Large Enough Rents Assumption always holds as long as $\lambda < 0.13$. This means that the expected time until the creation of the fund rewarding countries for pricing oil combustion is at least 7.54 years. It seems plausible that this holds. In 2010, the industrialized countries formulated the pledge to raise \$100billion per year for global environmental institutions. Policy-based reward funds such as the carbon pricing reward funds discussed here have been extensively discussed in the literature (e.g. in Cramton & Stoft (2012)). So arguably, as of 2021, at least 11 years have elapsed since the possibility of carbon pricing reward funds being created became salient.

6.2.3 Implications

According to the calibration presented above, the perverse incentive effect that arises before the global institution's creation cancels 62.5% of the welfare gains accruing during its existence, as the calibration presented here implies. Other types of global reward payment funds might be afflicted with similarly large perverse incentive effects.

For example, consider the case of a Global Health Security Challenge Fund that would reward countries based on their scores on the Global Health Security Index, as proposed by CSIS et al. (2020). Such an institution could

reduce the risk of future pandemics. Stern (2023) develops a static model in which countries choose the amount of effort they exert to prevent pandemics. The global institution's objective is to maximise the aggregate GHS Index score. Formally, the model is isomorphic to the static model of the current paper. As a result, in the dynamic extension as presented in section 6.1 the same results hold about the relative size of the perverse incentive effects. It is given by the same formula as in Lemma 11: $\frac{\delta+\kappa}{r+\kappa}$.

Let us now do an illustrative calibration. The global institution's discount rate δ is supposed to only be due to some exogenous events that would make it obsolete. According to the "time of perils hypothesis", humanity will eventually achieve existential security, for example through the future development and deployment of technologies that could cheaply protect against pandemics (Ord (2020), Aschenbrenner (2020)). This could then render the above-mentioned Global Health Security Challenge Fund obsolete. Suppose that the likely time for this to happen is in 50 years. This would suggest that $\delta = 1/50 = 0.02$. Assuming an interest as above $r = 0.05$, $\kappa = 0.08\bar{3}$, we find by Lemma 11 that in a world were global institutions condition on countries' past actions, the resulting perverse incentive effects before the global institution's creation cancel $\frac{\delta+\kappa}{r+\kappa} = 77.5\%$ of the expected discounted welfare gains realized during its existence.

It thus seems that the perverse incentive effects are likely to loom large for many of the important potential future reward payment funds. This suggests that it might be beneficial for the world to settle on a norm requiring newly created global institutions tasked with rewarding countries for globally beneficial policies (e.g. carbon taxes, spending on disease surveillance, etc.) to abstain from conditioning their reward payments on countries' past actions and instead aim to implement the optimal anonymous mechanism conditioning reward payments only on the current policy choices.

It might be objected that the world should instead settle on the following alternative norm: "GPGIs should never condition reward payments on countries' policies at times shortly before their creation". So for example, the GPGIs proposed in this paper would be allowed to condition reward payment to each country on its tax rate 20 years before the GPGI's creation (e.g. by rewarding countries based on the increase in the tax rate relative to the point in time 20 years before the GPGI's creation). Such a norm could mitigate the perverse incentives whilst still allowing the global institution to reduce informational rents. The extent to which informational rents would actually be reduced depends on how much types evolve over time. This could be quantified using the a available data on past fossil fuel tax rates.

A potential downside of such a more permissive norm might be that there could be a slippery slope to abandoning it. In some potentially very important future GPGIs, reward payments would be conditioned on newly introduced policies. For example, metagenomic sequencing might be essential for disease surveillance to detect novel pathogens Consortium, T. N. A. O. (2021). By the time that a GPGI might be created to reward countries on the basis of their spending on metagenomic disease surveillance, countries might only very recently have started to do such surveillance. There might then be a temptation for the GPGI to condition reward payments on increases in spending relative to the spending just before the GPGI's creation. But the expectation of this would create the perverse incentives we have analyzed in this section. A clear norm against all forms of conditioning on a country's past actions could avoid this.

7 Conclusion and limitations

This paper proposes a simple contract theory model. It provides simple explicit formulae for the optimal way for a global institution to reward countries for taxing their fossil fuels. Based on it, I have proposed a procedure for determining a reward payment function in practice.

A major limitation of the procedure is that it does not allow for learning over time. If and when the mechanism is implemented, we can observe how countries actually respond to the reward payment functions. It is unlikely that this will correspond very closely to the prediction of the model. It seems important to develop more flexible

contract theory models for constructing procedures for adjusting the reward payment function over time in light of the observed changes in tax rates.

For this, it would be important to extend the model by having the countries' types stochastically evolve over time and calibrate the extent of such changes based on countries' past paths of fossil fuel tax rates. Such a model could allow for a better assessment of whether global institutions should condition reward payments on countries' choices prior to their creation or not. If the types change a lot over time then mechanisms well-tailored to past types might perform poorly later on when the types will have changed. This would strengthen the case for not conditioning on past choices.

A model with stochastically evolving types could also help answer the question as to whether it would be innocuous for the global institution to use observed past choices to continually update its belief about the distribution of types. A country's past actions would then indirectly influence its reward payments in the present. For small countries this seems innocuous, since such countries would have only a small effect on the estimated distribution of types. However, for larger countries such as China and India this might be different. Quantifying these effects seems important.

Another priority should be to move beyond the casual empiricism of this paper's calibration and to carefully estimate the actual distribution of current tax rates on different fossil fuels, differentiated by narrow categories of fossil use (e.g. "oil used for heating", "gasoline", etc.).

Moreover, to properly assess the optimal roles to be played by carbon pricing reward funds on the one hand and project-based contracting approaches to mitigation on the other hand, it would be important to distill from the large literature on existing project based schemes like the Clean Development Mechanism best estimates about what fraction of projects are actually additional. The optimistic model used in this paper for analyzing project based contracting assumes that all projects are actually additional. Thus a more careful analysis will lead to lower performance estimates for project based contracting. This will strengthen the case for creating carbon pricing reward funds relative to what this paper's results suggest.

A Proofs for the static model

A.1 Proof of Lemma 3

Proof. Let L denote the global institution's Lagrangian and λ the Lagrange multiplier associated with the budget constraint:

$$L = \int_{\gamma=\underline{\gamma}}^{\bar{\gamma}} f(\gamma)(v(\tau(\gamma),\gamma) - \eta y(\tau(\gamma)) - \lambda t(\tau(\gamma)))d\gamma$$

From now on let us focus on the case where M is sufficiently small for the voluntary participation constraints to all be binding. We get:

$t(\tau(\gamma)) = -v(\tau(\gamma),\gamma)$ and thus:

$$L = \int_{\gamma=\underline{\gamma}}^{\bar{\gamma}} f(\gamma)(v(\tau(\gamma),\gamma) - \eta y(\tau(\gamma)) + \lambda v(\tau(\gamma)))d\gamma$$

The first order condition is:

$$\frac{\partial v}{\partial \tau}(\tau(\gamma),\gamma) - \eta y'(\tau) + \lambda \frac{\partial v}{\partial \tau}(\tau(\gamma),\gamma) = 0$$

Now using that $y(\tau) = -pe_d - e_d\tau + e_d + 1$ and $v(\tau(\gamma),\gamma) = -\frac{1}{2}e_d(\tau-\gamma)^2$, we obtain:

$$-(\lambda+1)e_d(\tau-\gamma)+\eta e_d=0$$

Thus the optimal contract is given by:

$$\tau(\gamma)=\gamma+\frac{\eta}{\lambda+1}$$

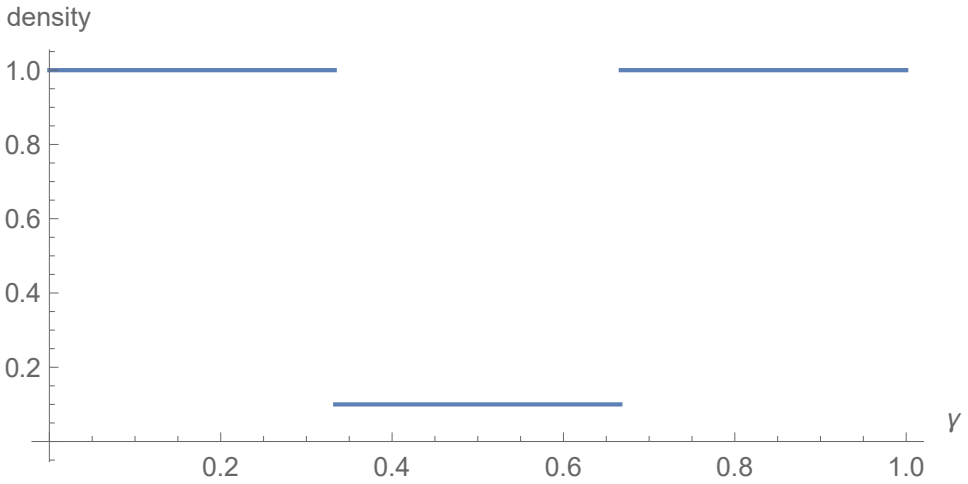
Integrating yields:

$$M=\int_{\gamma=\underline{\gamma}}^{\bar{\gamma}} t(\gamma)d\gamma=\frac{e_d}{2}(\frac{\eta}{\lambda+1})^2$$

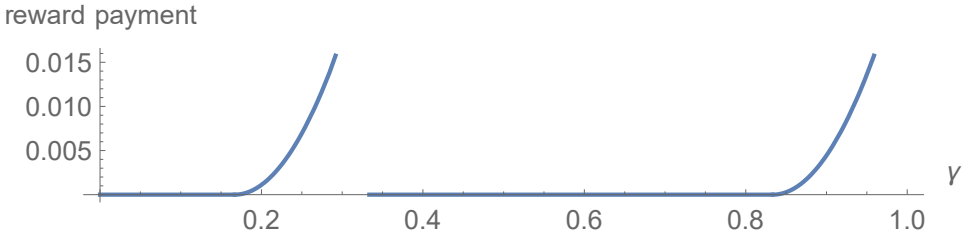
By eliminating the Lagrange multiplier λ, we obtain the claimed results. □

A.2 Proof of Lemma 4

Remark 1. Under complete information, it is clear that the optimal mechanism never involves inducing any type γ to choose a tax rate $\tau(\gamma)$ strictly below its preferred value. One might be inclined to think that this will also always be the case under incomplete information. However, this is not quite true. To see why, consider a type distribution distribution of the following form:



If the density in the middle part is sufficiently small, then it the decisions that they will be induced to make will become of neglible importance. Now consider the case where the global institution has a small budget. Then the optimal reward payment function looks like this:



This is because, by assumption, the density in the middle part of the distribution is so small that we can ignore it. Moreover, also by assumption, the budget at the global institution's disposal is so small that there is no concern that the types at the upper third of the distribution will be tempted to choose a tax rate intended for the types at the lower third of the distribution (i.e. only local incentive compatility constraints will bind for the upper third of the distribution). Then the problem reduces to solving two separate problems (with a common budget constraint): one for the upper third of the type space and one for the lower third of the type space. These problems are solved by parabolic reward payment functions tangent to the x-axis as illustrated in the above plot and as proved in lemma 7.

Now consider the types in the middle third of the distribution. The very lowest of them will be induced to increase their tax rates. But the types just above where the reward payment function reverts to 0 will find it optimal to distort downwards her tax rate.

This example shows that we need to make some assumption on the type distribution. In the lemma, we make the assumption that the density function be non-decreasing, which rules out the above example. Hopefully, this example will kindle the reader's interest for the following proof:

Proof. For a contradiction, suppose that at the optimal allocation there is some γ with $\tau(\gamma) < \gamma$. Consider the set of intervals containing only such values of γ. Moreover, restrict now attention to only those such intervals that are such that for no value of γ to the left of them we have $\tau(\gamma) < \gamma$. Formally, the interval we have thus define can be written as: $I = \cup\{[a, b] : \tau(c) < c \forall c \in [a, b] \, and \, \tau(c) \geq c \forall c < a\}$, where the set union is over all the intervals in the set defined.

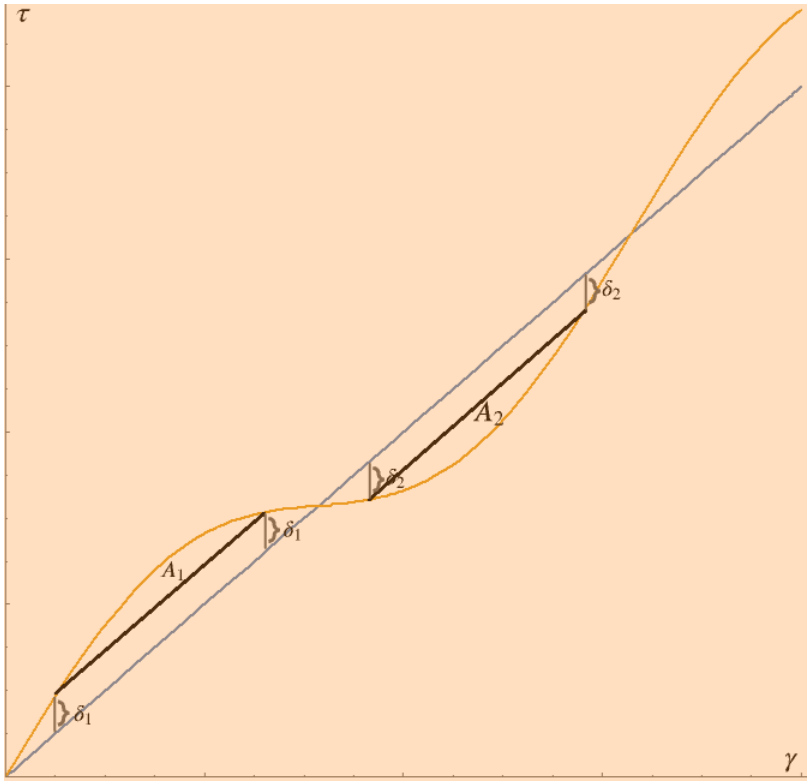
Now we can distinguish two cases:

Case 1: I contains the minimal γ in the support of the type space.

In this case, the participation constraint cannot be binding for any $\gamma < \sup I$, since the rents must be strictly decreasing on I, given that $U'(\gamma) = e_d(\tau(\gamma) - \gamma)$. Therefore we could then modify the allocation, setting $\tau(\gamma) = \gamma$ for all the types in I instead. This would decrease emissions, decrease countries' costs, whilst decreasing rents. Hence this would increase the Lagrangian, contradicting our assumption that we are dealing with an optimal mechanism. This case can therefore be ruled out.

Case 2: I does not contain the minimal γ in the support of the type space.

Hence the interesting case is where on some interval including the lowest type we have $\tau(\gamma) \geq \gamma$. An example of this case is depicted as the orange curve in the following diagram, where the blue curve is the identity function:



Now consider the following class of perturbations, indexed by time t: The allocation at time t is defined by the black lines on the corresponding part of the type space and by the orange line at all other types. The black lines are parallel to the blue identity function line and the regions A_1 and A_2 enclosed by the orange and the black lines each have area t. Let us denote by $\tilde{f}_1(t)$ and $\tilde{f}_2(t)$ the average density on the parts of the type space corresponding to the black lines.

Let us now consider the effect on the Lagrangian that arises from increasing t by a small amount dt. There are three effects: on rents, on aggregate emissions, and on countries' costs due to deviating from their otherwise preferred tax rate.

Consider first the effect on rents. Since $U'(\gamma) = e_d(\tau(\gamma) - \gamma)$, we can for each type γ compute her rent as the "aggregate signed area" enclosed between the orange and the blue line up to γ, where the area below the blue line has negative sign and the area above it positive sign. From this it follows that, by construction, our perturbation leaves unchanged the rents of the types to the right of the right black bar. For the types below this, rents can only decrease.

Given that emissions are linear in the tax rate, specifically $y(\tau) = 1 - e_d\tau$, the effect on emissions is $\tilde{f}_1(t)dt - \tilde{f}_2(t)dt$. Since we are assuming that the density is non-decreasing, we have that $\tilde{f}_2(t) \geq \tilde{f}_1(t)dt$, which implies that emission cannot increase.

Given that type γ's cost of deviating from choosing $\tau = \gamma$ is

$$-v(\tau, \gamma) := \frac{1}{2} e_d(\tau - \gamma)^2$$

we have

$$\frac{d}{d\tau}(-v(\tau,\gamma)) = -e_d(\tau-\gamma)$$

Hence the aggregate change in the countries' costs from deviating from their preferred tax rate is:

$$-\tilde{f}_1\delta_1 e_d - \tilde{f}_2\delta_2 e_d$$

Thus we always have a reduction in these costs.

Overall, we thus have: countries' costs are decreased, rents are decreased and emissions are decreased. This unambiguously means an increas in the value of the Lagrangian. Hence the curve in the diagram cannot correspond to an optimal allocation. □

A.3 Proof of Lemma 5

Proof. Let L denote the global institution's Lagrangian and λ the Lagrange multiplier associated with the budget constraint:

$$L = \int_{\gamma=\underline{\gamma}}^{\bar{\gamma}} f(\gamma)(v(\tau(\gamma),\gamma) - y(\tau)\eta - \lambda(-v(\tau(\gamma),\gamma) + U(\gamma)))d\gamma$$

$$L = \int_{\gamma=\underline{\gamma}}^{\bar{\gamma}} f(\gamma)(-y(\tau)\eta - \lambda U(\gamma) + (\lambda+1)v(\tau(\gamma),\gamma)) \tag{5}$$

Using integration by parts, we get:

$$\int_{\gamma=\underline{\gamma}}^{\bar{\gamma}} f(\gamma)U(\gamma)d\gamma = [-(1-F(\gamma))U(\gamma)]_{\gamma=\underline{\gamma}}^{\bar{\gamma}} + \int_{\gamma=\underline{\gamma}}^{\bar{\gamma}} (1-F(\gamma))U'(\gamma)d\gamma$$

$$\int_{\gamma=\underline{\gamma}}^{\bar{\gamma}} f(\gamma)U(\gamma)d\gamma = U(\underline{\gamma}) + \int_{\gamma=\underline{\gamma}}^{\bar{\gamma}} (1-F(\gamma))U'(\gamma)d\gamma$$

Now using that $U'(\gamma) = y(\gamma) - y(\tau(\gamma))$, we get:

$$\int_{\gamma=\underline{\gamma}}^{\bar{\gamma}} f(\gamma)U(\gamma)d\gamma = U(\underline{\gamma}) + \int_{\gamma=\underline{\gamma}}^{\bar{\gamma}} (1-F(\gamma))(y(\gamma) - y(\tau(\gamma)))d\gamma$$

Substituting this into equation 5 yields:

$$L = \lambda U(\underline{\gamma}) + \int_{\gamma=\underline{\gamma}}^{\bar{\gamma}} f(\gamma)(-y(\tau)\eta - \lambda\frac{1-F(\gamma)}{f(\gamma)}(y(\gamma) - y(\tau(\gamma))) + (1+\lambda)v(\tau(\gamma),\gamma))$$

The first order condition for τ is thus:

$$-y'(\tau)\eta + \lambda\frac{1-F(\gamma)}{f(\gamma)}y'(\tau(\gamma)) + (1+\lambda)\frac{\partial v}{\partial \tau}(\tau,\gamma) = 0$$

Now using that $y'(\tau) = -e_d$ and $\frac{\partial v}{\partial \tau}(\tau,\gamma) = e_d(\gamma-\tau)$, we obtain the claimed formula:

$$\eta - \lambda\frac{1-F(\gamma)}{f(\gamma)} + (1+\lambda)(\gamma-\tau) = 0$$

□

A.4 Proof of Lemma 6

Proof. The first order condition from Lemma 5 yields:

$$\tau = \frac{\eta}{\lambda+1} - \frac{\lambda}{\lambda+1}\frac{1-F(\gamma)}{f(\gamma)} + \gamma$$

Now using the uniform distribution assumption 2 yields:

$$\tau = \frac{\eta}{\lambda+1} - \frac{\lambda}{\lambda+1}(\bar{\gamma}-\gamma) + \gamma$$

Using the definition of μ, the first order condition becomes:

$$\tau = \mu\eta - (1-\mu)(\bar{\gamma}-\gamma) + \gamma$$

Recall that the first order condition is only relevant for $\gamma \geq \gamma_L$. The types $\gamma < \gamma_L$ do not participate and thus choose $\tau = \gamma$. Thus the formula for τ holds.

Now recall that there exists γ_L such that all $\gamma < \gamma_L$ do not participate whilst all $\gamma > \gamma_L$ do participate. We can distinguish the following two cases:

case 1: $\gamma_L > \underline{\gamma}$

By the formula for the optimal $\tau(\gamma)$, this case corresponds to $\mu < \frac{\bar{\gamma}-\underline{\gamma}}{\eta+\bar{\gamma}-\underline{\gamma}}$ and in this case we have:

$$\gamma_L = \bar{\gamma} - \eta\frac{\mu}{1-\mu}$$

The budget that needs to be spent is obtained by integrating the transfers, yielding:

$$m(\mu) := \frac{e_d\eta^3(2-\mu)\mu^3}{6(1-\mu)^2(\bar{\gamma}-\underline{\gamma})}$$

$m(\mu)$ is strictly increasing on $[0,1]$, since $m'(\mu) = \frac{e_d\eta^3\mu^2(3-(3-\mu)\mu)}{3(1-\mu)^3(\bar{\gamma}-\underline{\gamma})}$.

Moreover:

$$m(0) = 0$$

$$m(\frac{\bar{\gamma}-\underline{\gamma}}{\eta+\bar{\gamma}-\underline{\gamma}}) = \frac{e_d\eta(\bar{\gamma}-\underline{\gamma})^2(2\eta+\bar{\gamma}-\underline{\gamma})}{6(\eta+\bar{\gamma}-\underline{\gamma})^2}$$

This shows that the restriction of $\mu(M)$ to $[0, \frac{e_d\eta(\bar{\gamma}-\underline{\gamma})^2(2\eta+\bar{\gamma}-\underline{\gamma})}{6(\eta+\bar{\gamma}-\underline{\gamma})^2}]$ is the unique inverse of the restriction of $m(\mu)$ to $[0, \frac{\bar{\gamma}-\underline{\gamma}}{\eta+\bar{\gamma}-\underline{\gamma}}]$.

case 2: $\gamma_L = \underline{\gamma}$

In this case, integrating the transfers yields:

$$\frac{e_d}{6}\left(\mu^2\left(3\eta^2 + 3\eta(\bar{\gamma}-\underline{\gamma}) + (\bar{\gamma}-\underline{\gamma})^2\right) - (\bar{\gamma}-\underline{\gamma})^2\right)$$

This is strictly increasing in μ. □

A.5 Proof of Lemma 7

Proof. Consider first a reward payment scheme of the following form:

$$\tilde{t}(\tau) = Max(0, R(a\tau + b\tau^2 + c))$$

where R denotes the operator that preserves the right half (relative to its extremum) of the parabola $a\tau + b\tau^2 + c$ but outputs 0 on the left half.

We will show that there exist unique values for a, b, c as a function of $\mu(M)$ such that $\tilde{t}$ implements the optimal schedule of tax rates, which by Lemma 6 we know to be $\tau(\gamma) = max(\gamma, \mu(M)\eta - (1 - \mu(M))(\bar{\gamma} - \gamma) + \gamma)$. We will then show that the corresponding α, β, τ^* are as claimed.

Consider a type γ and suppose that a, b, c are such that γ participates when faced with $\tilde{t}$, in the sense of optimally choosing a τ with $\tilde{t}(\tau) > 0$. The type τ has utility $v(\tau, \gamma) + \tilde{t}(t)$. By Lemma 2 we have $v(\tau, \gamma) = -\frac{1}{2}e_d(\tau - \gamma)^2$, so the first order condition for γ's choice of τ becomes:

$$e_d(\tau - \gamma) = a + 2b\tau$$

Solving for τ gives:

$$\tau = \frac{a + e_d\gamma}{e_d + 2b} \tag{6}$$

By Lemma 6, we know that for the participating types, the optimal τ is given by:

$$\tau = \mu\eta - (1 - \mu)(\bar{\gamma} - \gamma) + \gamma \tag{7}$$

The functions defined by equations 6 and 7 are identical iff the following holds:

$$a = \frac{e_d\eta\mu - (1 - \mu)e_d\bar{\gamma}}{2 - \mu} \tag{8}$$

$$b = \frac{(1 - \mu)}{2(2 - \mu)}e_d \tag{9}$$

With this, one finds that if a type γ participates, then his optimal choice is $\mu\eta - (1 - \mu)(\bar{\gamma} - \gamma) + \gamma$, which coincides exactly with the choice at the optimal mechanism by Lemma 6. To check this, one can first insert this expression and the expressions for a and b into the first order condition. The second order condition is:

$$-e_d + 2b\tau < 0 \tag{10}$$

Inserting b into this equation yields:

$$-e_d + 2\frac{(1 - \mu)}{2(2 - \mu)}e_d < 0 \tag{11}$$

which is equivalent to

$$\frac{-1}{(2 - \mu)} < 0 \tag{12}$$

which is indeed always satisfied.

Now we need to choose the constant c such that exactly the same types actually participate as at the optimal mechanism. Consider first the case with partial participation. By Lemma 6 we know that the minimal type that

participates is $\gamma_L = \bar{\gamma} - \eta\frac{\mu}{1-\mu}$ and this type is indifferent between participating or not. Moreover, this type chooses $\mu\eta - (1-\mu)(\bar{\gamma} - \gamma_L) + \gamma_L = \gamma_L$. Thus under the quadratic reward payment scheme $\tilde{t}$, the type γ_L would get the following utility if it were to participate:

$$a\gamma_L + b\gamma_L^2 + c = a(\bar{\gamma} - \eta\frac{\mu}{1-\mu}) + b(\bar{\gamma} - \eta\frac{\mu}{1-\mu})^2 + c$$

Thus in order for γ_L to also be indifferent between participating or not, we need precisely the following:

$$c = -a(\bar{\gamma} - \eta\frac{\mu}{1-\mu}) - b(\bar{\gamma} - \eta\frac{\mu}{1-\mu})^2$$

Substituting in the expression 8 for a and expression 9 for b yields:

$$c = -\frac{e_d(\eta\mu + \bar{\gamma}(1-\mu))^2}{2(2-\mu)(1-\mu)}$$

Now let us verify for the case with partial participation that $\tilde{t}$ is identical to the $t(\tau) = \alpha(\tau - \tau^*) + \beta(max(0, \tau - \tau^*))^2$ as defined in the statement of the proposition. For $\tau \geq \tau^*$we simply need to verify that the coefficients of the quadratic coincide, i.e. that the following holds:

$$a = \alpha - 2\beta\tau^*$$

$$b = \beta$$

$$c = -\alpha\tau^* + \beta\tau^{*2}$$

The computations verifying these identities are omitted here.

Since $\alpha = 0$, it follows that $\tilde{t}$ also agrees with the claimed t on the remaining part of the domain, i.e. where $\tau < \tau^*$.

Now consider the case with full participation. In this case we need to set c such that the type γ_L gets 0 utility. This is achieved by:

$$c = -\frac{e_d\gamma_L^2(1-\mu)}{2(2-\mu)}$$

Again, we can verify the identity of $\tilde{t}$ with the t defined in the proposition by comparing the coefficients. □

A.6 Proof of Lemma 8

Proof. Consider a fixed M and let us vary η. We clearly must have $lim_{\eta\to\infty}\mu = 0$, since μ is one over the marginal value of relaxing the budget constraint. Now the formula holding under partial participation, $\frac{e_d\eta^3(2-\mu)\mu^3}{6(1-\mu)^2(\bar{\gamma}-\underline{\gamma})} = M$, implies that $lim_{\eta\to\infty}\eta\mu = (\frac{3(\bar{\gamma}-\underline{\gamma})}{e_d}M)^{\frac{1}{3}}$.

Hence the set of participating types converges to $[\bar{\gamma} - (\frac{3(\bar{\gamma}-\underline{\gamma})}{e_d}\mathrm{M})^{\frac{1}{3}}, \bar{\gamma}]$ and the tax rate converges to $\tau(\gamma) = \gamma + (\frac{3(\bar{\gamma}-\underline{\gamma})M}{e_d})^{\frac{1}{3}} - (\bar{\gamma} - \gamma)$. Moreover, by Lemma 7, the coefficients of the reward function become:

$\alpha = 0, \beta = \frac{1}{4}e_d, \tau^* = \bar{\gamma} - (\frac{3(\bar{\gamma}-\underline{\gamma})}{e_d}M)^{\frac{1}{3}}$

Now consider the case with full participation. We get:

$lim_{\eta\to\infty}\mu\eta = \sqrt{\frac{2M}{e_d} + \frac{1}{3}(\bar{\gamma} - \underline{\gamma})^2}$
Hence
$lim_{\eta\to\infty}\tau(\gamma) = lim_{\eta\to\infty}\mu\eta - (1-\mu)(\bar{\gamma} - \gamma) + \gamma = \gamma + \sqrt{2\frac{M}{e_d} + \frac{1}{3}(\bar{\gamma} - \underline{\gamma})^2} - (\bar{\gamma} - \gamma)$
The resulting emissions reductions are given by:
$\int_{\gamma=\underline{\gamma}}^{\bar{\gamma}} e_d(\tau(\gamma) - \gamma)f(\gamma)d\gamma = e_d(\sqrt{2\frac{M}{e_d} + \frac{1}{3}(\bar{\gamma} - \underline{\gamma})^2} - \frac{(\bar{\gamma}-\underline{\gamma})^2}{2})$
Moreover, by Lemma 7, the coefficients of the reward function become:
$\alpha = \frac{\sqrt{\frac{2M}{e_d} + \frac{1}{3}(\bar{\gamma}-\underline{\gamma})^2} - (\bar{\gamma}-\underline{\gamma})}{\sqrt{2}} e_d$, $\beta = \frac{1}{4}e_d$, $\tau^* = \bar{\gamma} - \sqrt{2}(\bar{\gamma} - \underline{\gamma}) + (1+\sqrt{2})\sqrt{\frac{2M}{e_d} + \frac{1}{3}(\bar{\gamma} - \underline{\gamma})^2}$ □

B Allowing for paternalism: The optimal mechanism when the global institution discards the tax aversion term when evaluating welfare

B.1 The global institution's objective function under paternalism

Consider a country's utility:

$$u(\tau, \theta, \phi) = B(y(\tau)) - py(\tau) - \theta y + \sigma y - \phi\tau$$

This utility is to be interpreted as corresponding to the country's revealed preference: The model assumes that the countries adjust their carbon prices so as to maximize this utility.

The paper has assumed that the global institution is non-paternalistic in the sense that it takes this utility to also be the country's welfare (not including global climate change damages). As discussed in footnote 5, this approach is questionable: The terms of trade effect σy is zero-sum in nature. Also, the tax aversion term $\phi\tau$ can plausibly be attributed to effects that do not reflect welfare, for example biased beliefs (Douenne and Fabre (2020)).

Let us now study the case where the global institution discards the tax aversion term $\phi\tau$ completely when it evaluates the countries' welfare. For tractability, I will now also assume that the local externality coefficient θ is the same for all countries. Thus countries only differ in the coeffecient σ of the terms of trade effect and the tax aversion coefficient ϕ. It turns out that the global institution's problem has a solution analogous to the one in the non-paternalistic case analyzed in the main part of the paper.

B.2 The optimal mechanism under complete information

Under complete information, the global institution's problem is to choose a carbon tax rate $\tau(\gamma)$ for each type γ to solve the following:

Problem 3.

$$\max_{\tau(.)} \int_{\gamma=\underline{\gamma}}^{\bar{\gamma}} f(\gamma)(B(y(\tau)) - py(\tau) - \theta y - \eta y(\tau(\gamma)))d\gamma$$

under the following 2 constraints:
1. exogenous budget:

$$\int_{\gamma=\underline{\gamma}}^{\bar{\gamma}} f(\gamma)t(\tau(\gamma), \gamma)d\gamma \leq M$$

where M denotes the global institution's exogenous budget.
2. voluntary participation:

$$v(\tau(\gamma),\gamma)+t(\tau(\gamma))\geq 0$$

Lemma 13. *Under complete information, the optimal contract implements the following profile of tax rates:*
For $M\leq \frac{e_d}{6(\bar{\gamma}-\underline{\gamma})}((\eta+\theta-\underline{\gamma})^3-(\eta+\theta-\bar{\gamma})^3)$*:*

$$\tau(\gamma)=\gamma+(\frac{6(\bar{\gamma}-\underline{\gamma})M}{e_d})^{\frac{1}{2}}\frac{\eta+\theta-\gamma}{((\eta+\theta-\underline{\gamma})^3-(\eta+\theta-\bar{\gamma})^3)^{\frac{1}{2}}}$$

The associated welfare gains w *are given by*

$$w=\sqrt{\frac{2M}{3e_d}}\sqrt{\frac{(\eta+\theta-\underline{\gamma})^3-(\eta+\theta-\bar{\gamma})^3}{\bar{\gamma}-\underline{\gamma}}}-M$$

The resulting reduction in the use of the fossil fuel is given by:

$$\sqrt{\frac{2M}{3e_d}}\sqrt{\frac{(\eta+\theta-\underline{\gamma})^3-(\eta+\theta-\bar{\gamma})^3}{\bar{\gamma}-\underline{\gamma}}}$$

For $M\geq \frac{e_d}{6(\bar{\gamma}-\underline{\gamma})}((\eta+\theta-\underline{\gamma})^3-(\eta+\theta-\bar{\gamma})^3)$*:*

$$\tau(\gamma)=\theta+\eta$$

The associated welfare gains w *are given by :*

$$w=\frac{e_d}{6}\frac{(\eta+\theta-\underline{\gamma})^3-(\eta+\theta-\bar{\gamma})^3}{\bar{\gamma}-\underline{\gamma}}$$

The resulting reduction in the use of the fossil fuel is given by:

$$\frac{e_d}{3}\frac{(\eta+\theta-\underline{\gamma})^3-(\eta+\theta-\bar{\gamma})^3}{\bar{\gamma}-\underline{\gamma}}$$

Proof. Let L denote the global institution's Lagrangian and λ the Lagrange multiplier associated with the budget constraint:

$$L=\int_{\gamma=\underline{\gamma}}^{\bar{\gamma}} f(\gamma)(v(\tau(\gamma),\gamma)+\phi\tau(\gamma)-\sigma y(\tau(\gamma))-\eta y(\tau(\gamma))-\lambda t(\tau(\gamma)))d\gamma$$

From now on let us focus on the case where M is sufficiently small for the voluntary participation constraints to all be binding. We get:
$t(\tau(\gamma))=-v(\tau(\gamma),\gamma)$ and thus:

$$L=\int_{\gamma=\underline{\gamma}}^{\bar{\gamma}} f(\gamma)(v(\tau(\gamma),\gamma)+\phi\tau(\gamma)-\sigma y(\tau(\gamma))-\eta y(\tau(\gamma))+\lambda v(\tau(\gamma)))d\gamma$$

The first order condition is:

$$\frac{\partial v}{\partial\tau}(\tau(\gamma),\gamma)+\phi+(\sigma+\eta)y'(\tau)+\lambda\frac{\partial v}{\partial\tau}(\tau(\gamma),\gamma)=0$$

Now using that $y(\tau) = -pe_d - e_d\tau + e_d + 1$ and $v(\tau(\gamma), \gamma) = -\frac{1}{2}e_d(\tau - \gamma)^2$, we obtain:

$$-(\lambda + 1)e_d(\tau - \gamma) + \phi + (\eta + \sigma)e_d = 0$$

Thus the optimal contract is given by:

$$\tau(\gamma) = \gamma + \frac{\eta + \sigma + \frac{\phi}{e_d}}{\lambda + 1}$$

Now using the definition $\gamma := \theta - \sigma - \frac{\phi}{e_d}$ yields:

$$\tau(\gamma) = \gamma + \frac{\eta + \theta - \gamma}{\lambda + 1}$$

Integrating yields:

$$M = \int_{\gamma=\underline{\gamma}}^{\bar{\gamma}} f(\gamma)t(\gamma)d\gamma = \int_{\gamma=\underline{\gamma}}^{\bar{\gamma}} \frac{e_d}{2(\bar{\gamma} - \underline{\gamma})}(\frac{\eta + \theta - \gamma}{\lambda + 1})^2 d\gamma = \frac{e_d}{6(\bar{\gamma} - \underline{\gamma})(\lambda + 1)^2}((\eta + \theta - \underline{\gamma})^3 - (\eta + \theta - \bar{\gamma})^3)$$

$$\lambda + 1 = (\frac{e_d}{6(\bar{\gamma} - \underline{\gamma})M})^{\frac{1}{2}}((\eta + \theta - \underline{\gamma})^3 - (\eta + \theta - \bar{\gamma})^3)^{\frac{1}{2}}$$

By eliminating the Lagrange multiplier λ, we obtain:

$$\tau(\gamma) = \gamma + (\frac{6(\bar{\gamma} - \underline{\gamma})M}{e_d})^{\frac{1}{2}} \frac{\eta + \theta - \gamma}{((\eta + \theta - \underline{\gamma})^3 - (\eta + \theta - \bar{\gamma})^3)^{\frac{1}{2}}}$$

Thus in particular if $M = \frac{e_d}{6(\bar{\gamma} - \underline{\gamma})}((\eta + \theta - \underline{\gamma})^3 - (\eta + \theta - \bar{\gamma})^3)$ then we get $\tau(\gamma) = \eta + \theta$ which is the first best under paternalism.

Now the welfare gains are:

$$w = \int_{\gamma=\underline{\gamma}}^{\bar{\gamma}} f(\gamma)(-\frac{1}{2}e_d((\frac{6(\gamma - \underline{\gamma})M}{e_d})^{\frac{1}{2}} \frac{\eta + \theta - \gamma}{((\eta + \theta - \underline{\gamma})^3 - (\eta + \theta - \bar{\gamma})^3)^{\frac{1}{2}}})^2$$

$$+e_d(\eta + \theta - \gamma)(\frac{6(\gamma - \underline{\gamma})M}{e_d})^{\frac{1}{2}} \frac{\eta + \theta - \gamma}{((\eta + \theta - \underline{\gamma})^3 - (\eta + \theta - \bar{\gamma})^3)^{\frac{1}{2}}})d\gamma$$

Integration yields:

$$w = \frac{1}{3}(\frac{6M}{e_d})^{\frac{1}{2}} \frac{((\eta + \theta - \underline{\gamma})^3 - (\eta + \theta - \bar{\gamma})^3)^{\frac{1}{2}}}{(\bar{\gamma} - \underline{\gamma})^{\frac{1}{2}}} - M$$

□

B.3 The optimal mechanism under incomplete information

Lemma 14. *Suppose that* $\frac{\eta+\phi}{\bar{\gamma}-\underline{\gamma}} > e_d$. *For any budget* $M \geq 0$, *let* $\mu(M) := \frac{1}{\lambda(M)+1}$, *where* $\lambda(M)$ *denotes the global institution's shadow value of funds when its budget is* M. *The optimal mechanism implements the following profile of tax rates:*

$$\tau(\gamma) = \gamma + max(0, \mu(\eta+\theta-\gamma) + (1-\mu)(\gamma-\bar{\gamma}))$$

Moreover, the following holds:
$\mu(0) = 0$
$\mu(M) = 1 \forall M \geq \frac{1}{2} e_d \eta(\eta + \bar{\gamma} - \underline{\gamma})$
$\mu(M)$ *is strictly increasing on* $[0, \frac{1}{2} e_d \eta(\eta+\bar{\gamma}-\underline{\gamma})]$
For $M \in [0, \frac{e_d(\bar{\gamma}-\underline{\gamma})^2(\eta+\theta-\underline{\gamma})(\eta+\theta-\bar{\gamma})}{3(\eta+\theta+\bar{\gamma}-2\underline{\gamma})^2}]$ *the* $\mu(M)$ *is given as the unique* $\mu \in [0, \frac{\bar{\gamma}-\underline{\gamma}}{\bar{\gamma}-\underline{\gamma}+\eta}]$ *satisfying:*

$$\frac{e_d(\eta+\theta-\bar{\gamma})^3(1-\mu)\mu^3}{3(1-2\mu)^2(\bar{\gamma}-\underline{\gamma})} = M$$

and the set of participating types is given by $[\bar{\gamma} - \eta\frac{\mu}{1-\mu}, \bar{\gamma}]$.
For $M \in [\frac{e_d(\bar{\gamma}-\underline{\gamma})^2(\eta+\theta-\underline{\gamma})(\eta+\theta-\bar{\gamma})}{3(\eta+\theta+\bar{\gamma}-2\underline{\gamma})^2}, \frac{1}{2} e_d(\eta+\theta-\underline{\gamma})^2]$ *the* $\mu(M)$ *is given as the unique* $\mu \in [0, \frac{\bar{\gamma}-\underline{\gamma}}{\bar{\gamma}-\underline{\gamma}+\eta}]$ *satisfying:*
$\frac{e_d}{6}\left(3\mu^2((\eta+\theta-\underline{\gamma})^2 - (\bar{\gamma}-\underline{\gamma})^2\right) = M$
and all types participate.
For $M > \frac{1}{2} e_d(\eta+\theta-\underline{\gamma})^2$ *we have* $\mu(M) = 1$ *and the first best is achieved.*

The optimal reward functions turn out to be flatter than in the non-paternalistic case. This can be explained as follows: The global institution is now also concerned about correcting the pre-existing inefficiencies due to the low (and often negative) fossil fuel tax rates. This weighs in favor of incentivizing the countries with low fossil fuel tax rates (including negative ones) to raise their fossil fuel tax rates.

C Proofs for the dynamic model

C.1 Proof of Lemma 11

Proof. Consider a point in time prior to the global institution's creation. Consider a country with an ex ante type γ deciding on its tax rate τ_1. (Since the probability rate for the global institution's creation is constant, the problem of choosing the tax rate τ_1 prior to the global institution's creation does not depend on calendar time.)

Let us denote by τ_2 the country's optimal choice of tax rate once the global institution exists. τ_2 is a random variable. It depends on the realization of the ex post type $\tilde{\gamma}$. Increasing τ_1 at the given point in time has the following downside for the country: If the global institution gets created at that point in time, then a higher τ_1 reduces the reward payment that the country will receive during the entire time that the global institution will exist. Since the probability rate of the global institution's creation is λ, this downside effect for the country of type γ is:
$-\lambda \int_{t=0}^{\infty} e^{-rt} e^{-\kappa t} t'(\tau_2 - \tau_1) dt = -\frac{\lambda}{r+\kappa} t'(\tau_2 - \tau_1)$
Thus the expected value of this effect for the type γ is:
$-\frac{\lambda}{r+\kappa} E(t'(\tau_2(\tilde{\gamma}, \tau_1) - \tau_1)|\gamma)$
Let us denote by $\tau_1(\gamma)$ the optimal choice for type γ. This is thus characterized by the following first order condition:

$$- e_d(\tau_1(\gamma) - \gamma) - \frac{\lambda}{r+\kappa} E(t'(\tau_2(\tilde{\gamma}, \tau_1) - \tau_1(\gamma))|\gamma) = 0 \tag{13}$$

The ex post first order condition characterizing $\tau_2(\tilde{\gamma}, \tau_1)$ is:

$$e_d(\tau_2(\tilde{\gamma}, \tau_1) - \tilde{\gamma}) = t'(\tau_2(\tilde{\gamma}, \tau_1) - \tau_1(\gamma))$$

Plugging this into the ex ante first order condition (equation 13) yields:

$$- e_d(\tau_1(\gamma) - \gamma) - \frac{\lambda}{r+\kappa} e_d E(\tau_2(\tilde{\gamma}, \tau_1(\gamma)) - \tilde{\gamma}|\gamma) = 0 \tag{14}$$

Now let us integrate this equation over the set of all types. Since we have normalised the measure of the set of all types to 1, we can view this integration as applying the expectation operator taken over γ:

$$- e_d E(\tau_1(\gamma) - \gamma) - \frac{\lambda}{r+\kappa} e_d E(E(\tau_2(\tilde{\gamma}, \tau_1(\gamma)) - \tilde{\gamma}|\gamma)) = 0 \tag{15}$$

By the law of iterated expectations, we have:
$E(E(\tau_2(\tilde{\gamma}, \tau_1(\gamma)) - \tilde{\gamma}|\gamma)) = E(\tau_2(\tilde{\gamma}, \tau_1(\gamma)) - \tilde{\gamma})$
Using this, we obtain:

$$- e_d E(\tau_1(\gamma) - \gamma) - \frac{\lambda}{r+\phi} e_d E(\tau_2(\tilde{\gamma}, \tau_1(\gamma)) - \tilde{\gamma}) = 0 \tag{16}$$

But by Lemma 1, $A_1(M) := \int -e_d(\tau_1(\gamma) - \gamma) f(\gamma) d\gamma = -e_d E(\tau_1(\gamma) - \gamma)$ is the aggregate increase in emissions in each of the periods prior to the global institution's creation. Similarly, $A_2(M) = \int_{(\gamma,\tilde{\gamma})} (e_d(\tau_2(\tilde{\gamma}, \tau_1(\gamma)) - \tilde{\gamma})) g(\gamma, \tilde{\gamma}) d\gamma d\tilde{\gamma} = E(\tau_2(\tilde{\gamma}, \tau_1(\gamma)) - \tilde{\gamma})$ is the decrease in emissions after the global institution's creation, where $g(\gamma, \tilde{\gamma})$ denotes the pdf of the joint distribution of γ and $\tilde{\gamma}$. We thus have:
$A_1(M) = \frac{\lambda}{r+\kappa} A_2(M)$
The existence of the mechanism has 2 effects on welfare: Those due to the induced change in tax rates prior to the creation of the global institution and those due to the induced change in tax rates during the global institution's existence.

Let us denote by w_1 the expected discounted change in welfare due to the first effect and by w_2 the expected discounted change in welfare due to the second effect. Since $e^{-\lambda t}$ is the probability that the global institution has not yet been created, we thus have:

$$w_1 = A_1(M) \int_{t=0}^{\infty} e^{-\delta t} e^{-\lambda t} dt = -\frac{\lambda}{\delta + \lambda} A_2(M)$$

Now consider the time during the global institution's existence. Since λ is the probability rate of the global institution's creation and since $e^{-\kappa(s-t)}$ is the probability that the global institution still exists at time s if it was created at time t, we have:

$$w_2 = A_2(M) \int_{t=0}^{\infty} e^{-\delta t} e^{-\lambda t} \lambda \int_{s=t}^{\infty} e^{-(\delta+\kappa)(s-t)} ds\, dt = A_2(M) \frac{\lambda}{(\delta + \lambda)(\delta + \kappa)}$$

Now using $A_1(M) = \frac{\lambda}{r+\kappa} A_2(M)$, we obtain for the ratio of negative perverse incentive effect before the global institution's creation and the positive effect during its existence:

$$\frac{-w_1}{w_2} = \frac{\delta + \kappa}{r + \kappa}$$

□

C.2 Proof of Lemma 12

Proof. Let us begin by showing that if the countries expect the global institution to base reward payments entirely on the change in tax rate relative to the moment just before its creation then the countries' optimal choices of tax rates before the global institution's creation are such that differences in tax rates equal differences in types.

Consider a point in time prior to the global institution's creation. Consider a country with an ex ante type γ deciding on its tax rate τ_1. (Since the probability rate for the global institution's creation is constant, the problem of choosing the tax rate τ_1 prior to the global institution's creation does not depend on calendar time.)

Let us denote by τ_2 the country's optimal choice of tax rate once the global institution exists. τ_2 is a random variable. It depends on the realization of the ex post type $\tilde{\gamma}$. Increasing τ_1 at the given point in time has the following downside for the country: If the global institution gets created at that point in time, then a higher τ_1 reduces the reward payment that the country will receive during the entire time that the global institution will exist. Since the probability rate of the global institution's creation is λ, this downside effect for the country of type γ is:

$-\lambda \int_{t=0}^{\infty} e^{-rt} e^{-\kappa t} t'(\tau_2 - \tau_1) dt = -\frac{\lambda}{r+\kappa} t'(\tau_2 - \tau_1)$

Thus the expected value of this effect for the type γ is:

$-\frac{\lambda}{r+\kappa} E(t'(\tau_2(\tilde{\gamma}, \tau_1) - \tau_1)|\gamma)$

Let us denote by $\tau_1(\gamma)$ the optimal choice for type γ. This is thus characterized by the following first order condition:

$$- e_d(\tau_1(\gamma) - \gamma) - \frac{\lambda}{r+\kappa} E(t'(\tau_2(\tilde{\gamma}, \tau_1) - \tau_1(\gamma))|\gamma) = 0 \tag{17}$$

The ex post first order condition characterizing $\tau_2(\tilde{\gamma}, \tau_1)$ is:

$$e_d(\tau_2(\tilde{\gamma}, \tau_1) - \tilde{\gamma}) = t'(\tau_2(\tilde{\gamma}, \tau_1) - \tau_1(\gamma))$$

Now recall that $\tilde{\gamma} = \gamma + \gamma^{\#}$ where the conditional distribution of $\gamma^{\#}$ given γ does not depend on γ. For each γ, we have the following system of equations:

$$- e_d(\tau_1(\gamma) - \gamma) - \frac{\lambda}{r+\kappa} E(t'(\tau_2(\tilde{\gamma}, \tau_1) - \tau_1(\gamma))|\gamma) = 0 \tag{18}$$

$$e_d(\tau_2(\tilde{\gamma}, \tau_1) - \gamma - \gamma^{\#}) = t'(\tau_2(\tilde{\gamma}, \tau_1) - \tau_1(\gamma))$$

Suppose we have a solution $(\tau_1(\gamma), \gamma^{\#} \mapsto \tau_2(\gamma + \gamma^{\#}, \tau_1))$ to this for some γ. Now consider a different type, denoted by γ'. Let us now check that the following is a solution of the system for γ':

$$(\tau_1(\gamma) + \gamma' - \gamma, \gamma^{\#} \mapsto \tau_2(\gamma + \gamma^{\#}, \tau_1) + \gamma' - \gamma)$$

To do this, we plug this into the focs for γ':

$$- e_d(\tau_1(\gamma) + \gamma' - \gamma - \gamma') - \frac{\lambda}{r+\kappa} E(t'(\tau_2(\gamma + \gamma^{\#}, \tau_1) + \gamma' - \gamma - (\tau_1(\gamma) + \gamma' - \gamma))|\gamma') = 0 \tag{19}$$

$$e_d(\tau_2(\gamma + \gamma^{\#}, \tau_1) + \gamma' - \gamma - \gamma' - \gamma^{\#}) = t'(\tau_2(\gamma + \gamma^{\#}, \tau_1) + \gamma' - \gamma - (\tau_1(\gamma) + \gamma' - \gamma))$$

This simplifies to the following:

$$- e_d(\tau_1(\gamma) - \gamma) - \frac{\lambda}{r+\kappa} E(t'(\tau_2(\gamma + \gamma^{\#}, \tau_1) - \tau_1(\gamma))|\gamma') = 0 \tag{20}$$

$$e_d(\tau_2(\tilde{\gamma}, \tau_1) - \gamma - \gamma^{\#}) = t'(\tau_2(\tilde{\gamma}, \tau_1) - \tau_1(\gamma))$$

But by our assumption that the conditional distribution of $\gamma^{\#}$ given γ' is the same as the conditional distribution of $\gamma^{\#}$ given γ this system is equal to the system that we had for γ, which by assumption is satisfied.

Thus we have established that if the global institution bases its reward payments on the change in tax rates then the differences in countries' tax rates prior to the global institution's creation equal the differences in their respective types. In other words: the fact that the global institution could be created at any time causes all countries to decrease their tax rate by an equal amount relative to what they would set in the absence of any mechanism.

Given that the component $\gamma^{\#}$ of the ex post type is distributed independently of the ex ante type γ, this implies that the ex post optimal mechanism must condition only on the change in tax rates. In fact, once the ex ante type is known, the global institution's ex post problem simply reduces to that in the static model.

Now what needs to be checked is whether countries are actually best off participating, as we have implicitly assumed in this proof so far, instead of just ignoring the mechanism altogether. By "participating" I will mean here "choosing tax rates such that one gets a strictly positive reward payments".

If a country participates in the mechanism then it experiences two effects on its utility: Those due to the induced change in tax rates prior to the creation of the global institution and those due to the rents it will earn during the global institution's existence. It follows from definition 3 that $R(M)$ is the aggregate amount of rents accruing to countries at each moment of the global institution's existence. Let u_2 denote countries' expected discounted utility gains due to these rents. We have:

$$u_2 = R(M) \int_{t=0}^{\infty} e^{-rt} e^{-\lambda t} \lambda \int_{s=t}^{\infty} e^{-(r+\kappa)(s-t)} ds\, dt = R(M) \frac{\lambda}{(r+\lambda)(r+\kappa)}$$

Analogously, let us denote by u_1 countries' expected discounted utility losses due to their decrease in tax rates before the global institution's existence. By Lemma 11, we know that the average decrease in tax rates before the global institution's creation equals $\frac{\lambda}{r+\kappa}$ times the average increase in tax rates during the global institution's existence. But now we also have that the aggregate decrease in tax rates before the global institution's creation is realized at minimal aggregate cost for the countries, namely by having each country reduce its tax rate by an identical amount. If the aggregate tax increases during the global institution's existence were also realized at minimal aggregate costs then the average cost incurred per time before the global institution's creation equals would be $(\frac{\lambda}{r+\kappa})^2$ times the average cost incurred by countries per time during the global institution's existence.

However, this is not the case for the increases in the tax rates caused by the global institution when it exists. In fact, the ex post optimal mechanism will always sacrifice some constrained efficiency (which would dictate uniform tax increases) for the sake of reducing informational rents. Hence we obtain that the aggregate cost incurred by countries per time before the global institution's creation is less than $(\frac{\lambda}{r+\kappa})^2(M - R(M))$. To see why, note that, by the definition of informational rents, $M - R(M)$ is the aggregate cost incurred by countries due to the increased tax rates. Hence we deduce:

$$u_1 < (\frac{\lambda}{r+\kappa})^2 (M - R(M)) \int_{t=0}^{\infty} e^{-rt} e^{-\lambda t} dt = \frac{\lambda^2}{(r+\kappa)^2(r+\lambda)} (M - R(M))$$

Countries are better off participating than not iff $u_1 < u_2$. Given the above results, for this it is sufficient that the following condition holds:

$$\frac{\lambda^2}{(r+\kappa)^2(r+\lambda)}(M - R(M)) < R(M)\frac{\lambda}{(r+\lambda)(r+\kappa)}$$

which simplifies to the following condition:

$$\frac{\lambda}{(r+\kappa)} < \frac{R(M)}{M - R(M)}$$

□

References

[1] Antón, A. (2020). Taxing crude oil: A financing alternative to mitigate climate change?. Energy Policy, 136, 111031.

[2] Armstrong, M., & Rochet, J. C. (1999). Multi-dimensional screening:: A user's guide. European Economic Review, 43(4-6), 959-979.

[3] Cramton, P., Ockenfels, A., & Stoft, S. (2017). 12 An International Carbon-Price Commitment Promotes Cooperation. Global Carbon Pricing, 221.

[4] Cramton, P., & Stoft, S. (2010). International climate games: From caps to cooperation. Available at SSRN 1646473.

[5] Cramton, P., & Stoft, S. (2012). Global climate games: How pricing and a green fund foster cooperation. Economics of Energy & Environmental Policy, 1(2), 125-136.

[6] Dahl, C. (2009). Energy demand and supply elasticities. Energy policy, 72.

[7] Dechezleprêtre, Antoine, Jonathan Colmer, Caterina Gennaioli, Matthieu Glachant, and Anna Schröder. Assessing the Additionality of the Clean Development Mechanism: Quasi-Experimental Evidence from India. Tech. rep, 2014.

[8] Douenne, T. (2020). The vertical and horizontal distributive effects of energy taxes: A case study of a french policy. The Energy Journal, 41(3).

[9] Freixas, X., Guesnerie, R., & Tirole, J. (1985). Planning under incomplete information and the ratchet effect. The review of economic studies, 52(2), 173-191.

[10] Gersbach, H., & Winkler, R. (2011). International emission permit markets with refunding. European Economic Review, 55(6), 759-773.

[11] Greaves, H. (2017). Discounting for public policy: a survey. Economics & Philosophy, 33(3), 391-439.

[12] Golombek, R., Irarrazabal, A. A., & Ma, L. (2018). OPEC's market power: An empirical dominant firm model for the oil market. Energy Economics, 70, 98-115.

[13] International Monetary Fund. (2019). Fiscal Policies for Paris Climate Strategies-from Principle to Practice.

[14] Laffont, J. J., & Martimort, D. (2009). The theory of incentives. Princeton university press.

[15] Laffont, J. J., & Tirole, J. (1993). A theory of incentives in procurement and regulation. MIT press.

[16] Martimort, D., & Sand-Zantman, W. (2016). A mechanism design approach to climate-change agreements. Journal of the European Economic Association, 14(3), 669-718.

[17] Martimort, D., & Sand-Zantman, W. (2013). Solving the global warming problem: beyond markets, simple mechanisms may help!. Canadian Journal of Economics/Revue canadienne d'économique, 46(2), 361-378.

[18] Ricke, Katharine, et al. "Country-level social cost of carbon." Nature Climate Change 8.10 (2018): 895-900.

[19] Roberts, J. T., Weikmans, R., Robinson, S. A., Ciplet, D., Khan, M., & Falzon, D. (2021). Rebooting a failed promise of climate finance. Nature Climate Change, 11(3), 180-182.

[20] Sovacool, B. K. (2017). Reviewing, reforming, and rethinking global energy subsidies: towards a political economy research agenda. Ecological Economics, 135, 150-163.

[21] Steckel, Jan Christoph, et al. "From climate finance toward sustainable development finance." Wiley Interdisciplinary Reviews: Climate Change 8.1 (2017): e437.

[22] Strand, J. (2020a). Supporting Carbon Tax Implementation in Developing Countries through Results-Based Payments for Emissions Reductions.

[23] Strand, J. (2020b). Transformational Climate Finance.