Social Planning and Coercion Under Bounded Rationality with an Application to Environmental Policy

Beat Hintermann (corresponding author)
University of Basel, Faculty of Business and Economics
Peter Merian-Weg 6
4002 Basel, Switzerland
b.hintermann@unibas.ch; +41 61 207 3339

Thomas F. Rutherford
University of Wisconsin at Madison
330 Taylor Hall
Madison, WI 53706
rutherford@aae.wisc.edu; +1 608 890 4576

Social Planning and Coercion Under Bounded Rationality: An Application to Environmental Policy

Abstract

We develop a theory of social planning with a concern for economic coercion, which we define as the difference between consumers' actual utility, and the "counterfactual" utility they expect to obtain if they were able to set policy themselves. Reasons to limit economic coercion include protecting minorities, preventing disenfranchised groups from engaging in socially costly behavior, or political economy considerations. If consumers are fully rational, we show that limiting coercion is equivalent to placing more welfare weight on coerced consumers at the expense of others. If, however, consumers' rationality is bounded, counterfactual utility becomes endogenous to current policy, and the welfare loss associated with limiting coercion increases. We set up a numerical version of our model and find that the bias-related welfare loss can be substantial.

1 Introduction

Group membership confers benefits. For example, being a resident of a community entails the right to access local public goods such as schools or parks; citizens of a nation state enjoy the benefits of defense and other national policies; and members of a supra-national union profit from improved terms of trade or a decrease in the probability of armed conflict. On the other hand, being part of a group also means being subject to its rules. Because it is unlikely that an individual agrees with every rule (or the degree to which they are applied) but is subject to these rules nonetheless, group membership implies a degree of coercion. In an economic context, coercion occurs wherever individuals are subject to decisions that affect their utility, but over which they have no control. Citizens of a nation state are coerced into accepting, for example, a particular level of public provision or redistribution of income that will generally differ from what they would prefer.[1]

Although economic coercion is a natural consequence of applying one set of rules to a heterogeneous population, it has been largely ignored in models of social planning. In this paper, we modify the standard social planner problem by incorporating a constraint on the maximally acceptable level of economic coercion that any consumer can be subjected to, and discuss the resulting implications for efficiency and welfare.

Defining coercion requires a reference to some counterfactual situation that differs from the status quo, and which may or may not be observable (Congleton, 2014). For example, the counterfactual implicit in a proposed sales contract is characterized by the seller turning over the good and the buyer releasing money. If the consumer has experience with the good in question and there is no private information, utility in the counterfactual situation is known to both involved parties. In contrast, when a discrete change in important policy variables is proposed (e.g., an environmental tax reform), the counterfactual could be a situation that has not been observed before, and which is therefore more difficult to assess by consumers.

We use as a starting point the framework developed by Winer et al. (2014), who are the first to introduce an explicit concern with coercion into a mainstream social planning context. We adopt their working definition of coercion as the difference between the

[1] Coercion is the "act, process or power of coercing", with the definition of the relevant verb given by Merriam-Webster as 1.) to restrain or dominate by force, 2.) to compel to an act or choice, or 3.) to achieve by force or threat. There are many forms of coercion. In this article, we abstract from physical coercion by assuming that the government has a mechanism to levy taxes, and focus on the coercion that arises from the necessity of using some sort of collective choice mechanism.

hypothetical utility a consumer thinks he should achieve, and the utility he obtains conditional on actual tax rates and public provision. Winer et al. (2014) focus on what they call the "individual-in-society" definition of coercion and assume that consumers arrive at their counterfactual by quantity-adjusting the level of the public good at a given tax price, whereas Sehili and Martinez-Vasquez (2014) use the dual approach of letting consumers choose the tax price at a given quantity of the public good. In our model, consumers treat neither the tax price nor the quantity of the public good as given, but compute their counterfactual utility by choosing all policy variables themselves.[2]

Gamson (1961) and Skaperdas (1992) discuss economic coercion during group formation, and Wallis (2014) and Skaperdas (2014) examine the role of coercion in emerging societies without a pre-existing rule of law.[3] In this paper, we focus on the situation where a group has already been formed and a democratic society is in place. In our example, the social planner (or equivalently, the government) decides about the level of public provision and environmental quality in a society where agents differ with respect to tastes and endowments. Unlike Ledyard (2014), we assume that individuals face prohibitively high material and emotional costs when leaving the group (e.g., the costs of migration and social adaptation). Instead of moving away, they are assumed to express their discontent by other means, for example engaging in labor strikes, mass protests, tax noncompliance or vandalism.

The economic subfield where coercion has traditionally been acknowledged is the normative theory of public choice, which is concerned with the definition of optimal decision rules for policy. Early contributers were Wicksell (1896) and Lindahl (1919), who proposed approximate unanimity as a direct consequence of the desire to minimize coercion, and Buchanan and Tullock (1962), who derive unanimity in the framing of a constitution as an efficiency condition. This literature challenges the traditional social planning paradigm in which the planner is allowed to coerce anyone as long as social welfare is improved, and social acceptance of government policy is simply assumed. By imposing a constraint on the maximally allowable level of coercion, we essentially adopt a Wicksellian view of public finance. As an alternative to our social planning framework, economic

[2]Winer et al. (2014) call this the "individual-as-dictator" approach. Note that we maintain the assumption that consumers take the given tax system as given in the sense that they can only adjust existing tax rates rather than overhaul the entire tax system.

[3]Even though members of a society are coerced by group rules, joining the group may convey greater utility for an individual than remaining outside, mainly due to the coercion of others, e.g., in the context of enforcing contracts (Baumol, 2004).

coercion could also be investigated based on a political economy framework, where a political candidate or a ruling party may want to limit coercion to some voter groups in order to maximize the chance of being elected.

Generally, all members of a group will be economically coerced to some extent, although the degree of coercion generally varies over the population. It can therefore not be the goal of government policy to eliminate coercion altogether, but there could be reasons to set a limit to the amount of coercion that any particular individual or group may be subjected to in order to reduce the possibility for social strife. Even though limiting coercion will reduce social welfare as traditionally defined, the resulting reduction in collateral damages and disruptions (which in our framework are not captured by the welfare function) may more than make up for the loss. Limiting coercion can therefore improve welfare in a more general sense.

An example of an explicit limit on coercion is the ”Takings Clause" of the Fifth Amendment of the U.S. Constitution, which limits the burden that can be placed on individuals in the context of financing public goods. It states that "private property shall not be taken for public use, without a just compensation". This means that the government cannot simply expropriate some individuals, even if society would benefit as a whole, and can be interpreted as a coercion limit for owners of land and capital.[4] On the other hand, there are F.D. Roosevelt's "Four Freedoms", one of which is the "Freedom from Want", which specifies that everyone has the right to basic necessities such as food, clothing and shelter, and which can be interpreted as a coercion limit for the poor and as a mandate for some expropriation of others.

Our paper is linked to the emerging literature of behavioral public economics, which allows for systematic departures from neoclassical principles.[5] We allow for the possibility that consumers make systematic mistakes when computing their counterfactual utility and investigate the consequences that this has for the resulting equilibrium and the associated levels of social welfare and coercion. Under full rationality and complete information, a concern with coercion can be addressed within the social welfare function by adding additional welfare weight to the most coerced groups. If, however, consumers make mistakes when computing their counterfactual utility, the latter becomes endogenous, and

[4]For a legal discussion, see e.g. http://law2.umkc.edu/faculty/projects/ftrials/conlaw/takings.htm, last accessed in August 2016.

[5]For a review of this literature, see Bernheim and Rangel (2007).

the social planner has to consider the change in counterfactual utility in response to a change in policy. We further show that the error in the computation of the counterfactual itself leads to a welfare loss, in addition to the unavoidable loss resulting from imposing a binding coercion constraint under the standard neoclassical assumptions. If counterfactual errors are sufficiently large and the coercion constraint sufficiently tight, the policy outcome may even be allocationally inefficient in the sense that it is inconsistent with social welfare maximization based on a set of non-negative welfare weights.

We operationalize our model numerically in a simplified setting with two policy dimensions and three consumer groups. Consumers derive utility from private consumption, a public good, and environmental quality (which is a type of public good over which tastes can differ significantly, for example in the context of climate change). The environment is adversely affected by the use of a polluting intermediate good, which can be imperfectly substituted with a clean input. We posit that consumers know the correct relationship between price changes and aggregate demand levels within markets at a given equilibrium, but that they neglect general equilibrium effects and treat the elasticity of labor supply as given. We find that even with this relatively mild departure from full rationality, the welfare loss that is attributable to the endogenous nature of the counterfactual is responsible for an important share of the total welfare loss that results from imposing a binding coercion constraint. This suggests that if the government wants to limit economic coercion to some consumer group, it is important to consider the consumer's view of the world, even (or especially) if this view is imperfectly informed. In contrast, ignoring the bias and setting policy as if consumers solved the full general equilibrium model correctly leads to an excess in economic coercion.

An important caveat is the high informational requirement implied by our model. The social planner not only needs to know private utilities (as in standard social planning), but also the type of error that consumers make in their computation of the counterfactual. This limits the application of our theory to contexts where information about this error exists or can be approximated.

The next section presents our theoretical framework and the implications of limiting economic coercion under limited rationality for welfare and efficiency. Section 3 presents our numerical version of the model and discusses the results, and Section 4 concludes.

2 Theoretical framework

We start with a basic social planning model and then explore the welfare effects of introducing a constraint on coercion.

2.1 Model

Let X_{hi} refer to the demand for (supply of) the private good (factor) $i = 1, ..., N$ by a consumer of type $h = 1, ..., H$. Consumers derive utility from private goods and factors as well as a public good G, which may be a vector of various publicly provided goods, and which may include environmental quality.[6]

Consumer h treats the level of public provision as given and maximizes his or her utility[7] by solving

$$\max_{X_{hi}} \quad U^h(X_{hi}; G) \qquad s.t. \qquad \sum_{i=1}^{N} P_i X_{hi} \leq I_h \tag{1}$$

where I_h is a fixed amount of (non-taxable) lump-sum income, and P_i is an element of the N-dimensional vector **P** referring to the consumer price of good/factor i, which may include a tax. The utility function is assumed to be twice differentiable and concave, and it satisfies the Inada conditions. Solving (1) gives rise to N demand and supply functions of the form $X_{hi} = X_{hi}(\mathbf{P}, I_h, G)$, and to the indirect utility function $V^h = V^h(\mathbf{P}, I_h, G)$.

The government funds the production of the public good by means of ad-valorem taxes on the private goods and factors, with good 1 serving as the numeraire:

$$P_i = p_i(1 + t_i) \quad \text{for} \quad i = 2, ..., N \tag{2}$$

Net-of-tax prices p_i are equal to the marginal rate of transformation between these goods (factors) and the numeraire. Tax revenue is fully used to finance the public good:[8]

[6] We need public provision in our model in order to justify the need for raising and spending government revenue. Furthermore, we need more than one public good to allow for disagreement (and thus scope for economic coercion) not only with respect to the size of the government, but also with respect to the composition of public expenditure. Since environmental quality, and especially climate change, is a particularly salient example of a public good over which disagreement exists, we include it in our numerical illustration below.

[7] To improve readability, this paper uses the masculine third-person personal pronoun in a gender-neutral sense for the remainder of the paper.

[8] We assume constant production costs of the public good and normalize its unit (or their units, if G is a vector) such that the cost is one. If public production technology is general rather than linear, the cost of public provision would change with a marginal change in tax rates along with private producer prices.

$$G = \sum_{i=2}^{N} p_i \, t_i \cdot \sum_{h=1}^{H} X_{hi}(\mathbf{P}, I_h, G) \tag{3}$$

Because the focus of the model is on consumers, we abstract from the production side by assuming production to be efficient. After further substituting the market-clearing condition for goods and factors, we express the production constraint as an aggregate production-possibilities frontier of the form

$$F\Bigg(\sum_{h=1}^{H} X_{hi}(\cdot); G\Bigg) \leq 0 \tag{4}$$

The government maximizes social welfare by choosing a vector of tax rates **t** subject to constraints on the technology and the budget requirement, and to an additional constraint on the maximum allowable level of economic coercion that a particular consumer type may experience.[9] From (2), producer prices are a function of tax rates and exogenous parameters representing consumers' tastes and the production technology, and we can therefore express indirect utility as a function of tax rates. We define economic coercion as the difference between actual welfare $V^h(\cdot)$ and "counterfactual" welfare $\tilde{V}^h(\cdot)$.

After substituting the government's budget constraint, the problem can be written as one where the social planner chooses a vector of tax rates **t** to maximize social welfare, subject to the technology and the coercion constraints:

$$\max_{\mathbf{t}} \quad W = \sum_{h=1}^{H} \alpha_h \cdot V^h\big(\mathbf{P(t)}; I_h; G(\mathbf{P(t)}, \mathbf{t}, \mathbf{I})\big) \tag{5}$$

$$s.t. \quad F\Bigg(\sum_{h=1}^{H} X_{hi}(\cdot); G(\cdot)\Bigg) \leq 0 \tag{6}$$

$$\tilde{V}^h(\cdot) - V^h(\cdot) \leq \bar{K}_h \quad \forall \; h \tag{7}$$

Individuals' utility enters social welfare according to the welfare weight α_h, with $\sum_{h=1}^{H} \alpha_h = 1$, and $\bar{K}_h$ is the limit of economic coercion which consumer h may experience.[10] This limit is assumed to be given, i.e. it is determined in the political process

[9] An alternative way to account for coercion would be to incorporate it into the welfare function, along with the potential consequences of civil unrest, rather than imposing an exogenously defined coercion constraint. However, this requires a (presumably exogenous) weighting of coercion relative to the utility from consumption, such that the fundamental problem of trading off welfare vs. coercion remains.

[10] We allow the coercion constraint to differ across consumer groups in order to avoid having to trade off one consumer's coercion against another's. Note that the coercion constraint will generally be binding for one group

along with the welfare weights.[11] The first-order necessary condition w.r.t. to tax $t_i \in \mathbf{t}$ is

$$\sum_{h=1}^{H}(\alpha_h + \kappa_h) \cdot \frac{\partial V^h(\cdot)}{\partial t_i} - \kappa_h \frac{\partial \tilde{V}^h(\cdot)}{\partial t_i} = \lambda \cdot \frac{\partial F(\cdot)}{\partial t_i} \tag{8}$$

where λ is the shadow price on the technology constraint, and κ_h are the shadow prices of the coercion constraints for each consumer type. The derivative of the indirect utility function with respect to taxes is the sum of the utility effects via prices, levels of demand and supply, and the public good. Denoting aggregate demand (supply) for good (factor) i over all households as $\sum_H X_{hi} \equiv X_i$ and suppressing function arguments, the marginal change in utility is given by

$$\frac{\partial V^h}{\partial t_i} = \frac{\partial V^h}{\partial P_i} p_i + \sum_{k=1}^{N} \frac{\partial V^h}{\partial P_k} \frac{\partial p_k}{\partial t_i}(1 + t_k) + \frac{\partial V^h}{\partial G} \frac{\partial G}{\partial t_i} \tag{9}$$

$$\frac{\partial G}{\partial t_i} = \frac{p_i X_i + \sum_{k=1}^{N} \frac{\partial p_k}{\partial t_i} t_k X_k + \sum_{k=1}^{N} p_k t_k \frac{\partial X_k}{\partial t_i}}{1 - \sum_{k=1}^{N} \left(\frac{\partial p_k}{\partial G} t_k X_k + \frac{\partial X_k}{\partial G} \right)} \tag{10}$$

The first term in (9) is the direct price effect on utility via expenditure on good X_{hi}. A marginal change in t_i also affects utility via changes in the producer price vector (second term) and the public good (third term). The quantity of public provision depends on t_i via the revenue generated in market i (the first term in the numerator of Eq. 10), net of revenue changes due to changes in prices (second term), and aggregate levels of demand and supply (third term). The denominator in (10) captures feedback effects that arise if a change in public provision affects the prices, and the demand and supply of private goods and factors.

The total marginal effect on the production side depends on aggregate levels of demand and supply, as well as on potential feedback effects if demand (supply) of private goods (factors) depends on the level of public provision:

$$\frac{\partial F}{\partial t_i} = \sum_{k=1}^{N} \frac{\partial F}{\partial X_k} \frac{\partial X_k}{\partial t_i} + \frac{\partial F}{\partial G} \frac{\partial G}{\partial t_i} \tag{11}$$

We now turn to counterfactual utility, which describes the utility the consumer believes

only, such that there is no qualitative difference between uniform and individual coercion constraints.

[11] The normative foundation for the coercion limit is implicit in the work by Wicksell (1896). If society accepts unlimited coercion in the interest of maximizing overall welfare, as defined by an individualistic welfare function, then the limit should be zero. If, however, society places a value on not inordinately coercing its members, then the limits to coercion have to be determined jointly with the weights in the social welfare function. Whether a democratic society is able to agree on such limits is, just like the social welfare function itself, subject to the critique raised by Arrow's (1951) impossibility theorem.

he obtains under his preferred policy. In the counterfactual problem, consumers choose consumption along with the vector of tax rates $\mathbf{t}^h$ (and thus the level of public provision via the budget constraint) that they prefer, subject to a technology constraint, which may or may not be the same as the one considered by the government.

Consumers can make a range of different errors when solving (12). In the most basic sense, they may fail to consider the budget and technology constraints. Even if consumers are fully informed and rational about budget balance and production possibilities, they may not foresee all direct and indirect effects on producer prices and demand levels that result from a change in tax rates. In our application, we assume that consumers neglect general equilibrium effects.[12] If consumers' computations involve a bias, the solution to this problem may depend on the current policy $\bar{\mathbf{t}}$. Intuitively, if the policy is changed from the benchmark B to some alternative policy point A, consumers realize that their counterfactual computation for their utility level at A was incorrect (due to the bias) and therefore update their expected utility associated with some other point A′. This means that the counterfactual utility under policy A′ is not the same if computed from B or from A, making the counterfactual endogenous to the current policy.[13]

The Lagrangian for consumer h's problem can be written as

$$\max_{X_{hi},\mathbf{t}^h} \quad \tilde{\mathscr{L}}^h = U^h\big(X_{hi};\tilde{G}(\mathbf{t}^h,\bar{\mathbf{t}}),\mathbf{I}\big) + \tilde{\mu}_h\Bigg(I_h - \sum_{i=1}^{N} P_i X_{hi}\Bigg) - \tilde{\lambda}_h\tilde{F}\Bigg(\sum_{k=1}^{H}\tilde{X}_{ki}(\tilde{\mathbf{P}}(\mathbf{t}^h,\bar{\mathbf{t}}),\mathbf{I});\tilde{G}(\mathbf{t}^h,\bar{\mathbf{t}}),\mathbf{I})\Bigg) \tag{12}$$

Solving (12) and substituting the derived counterfactual demand levels into the utility function yields an indirect counterfactual utility function that can be written as

$$\tilde{V}^h = \tilde{V}^h[\mathbf{I};\epsilon_h(\bar{\mathbf{t}})] \tag{13}$$

Counterfactual indirect utility depends on the exogenous parameters such as income and technology, and on the way consumers solve the problem, which we notationally

[12]Technically speaking, all derivatives in the summation terms of (8)-(11) that involve goods or factors $k \neq i$ are general equilibrium effects. Any or all of these may be ignored even by quite rational consumers. A different type of "error" could occur if consumers social welfare (as opposed to their private utility), as has been suggested by Boadway (2014).

[13]The update leads to an improvement, but not necessarily to a correction of the counterfactual utility at A′. Naturally, if no updating takes place, consumers' counterfactual (even if incorrect) does not depend on current policy. A biased computation of counterfactual utility is therefore a necessary, but not a sufficient condition for the counterfactual to be endogenous with current policy.

represent by the error $\epsilon_h(\bar{\mathbf{t}})$.

The source of the endogeneity of consumers' counterfactual utility, as well as its importance for the coercion-constrained equilibrium, can be seen more clearly by returning to the government's problem. Application of the envelope theorem implies that the derivative of the counterfactual utility $\tilde{V}^h$ with respect to t_i is equal to the derivative of the counterfactual Lagrangian, evaluated at the counterfactual demand levels and tax rates chosen by consumers. Dropping income levels as arguments for convenience, this can be written as

$$\frac{\partial \tilde{V}^h\big(\epsilon_h(\mathbf{t})\big)}{\partial t_i} = \frac{\partial \tilde{\mathscr{L}}^h\big(\epsilon_h(\mathbf{t})\big)}{\partial t_i}\bigg|_{\tilde{X}^*_{hi},\,\mathbf{t}^{h*}} \tag{14}$$

If consumers solve the counterfactual problem correctly (or at any rate, in the same way as the government), it follows that $\partial\tilde{\mathscr{L}}^h/\partial t_i = \partial\tilde{\mathscr{L}}^h/\partial t_i^h = 0\ \forall\ i$, and the last term on the left-hand side of (8) drops out.[14] In this case, the welfare weight for type h is simply increased from α_h to $\alpha_h+\kappa_h$. In other words, limiting coercion under full rationality places additional welfare weight on the consumer type(s) for which the coercion constraint is binding, but all other terms related to counterfactual utility drop out of the model. The reason is that correctly computed counterfactuals are a function only of fundamentals such as consumer preferences, technology and the availability of resources. Since these primitives are fixed, counterfactual utility becomes a constant, and like any constant it drops out of the first-order conditions.[15]

If, however, consumers make mistakes when computing the counterfactual, and these mistakes are sensitive to marginal policy choices, then (14) is not zero, and consumers' counterfactuals are endogenous to policy. The coercion-concerned social planner needs to take the change in counterfactual utility in response to a change in a policy variable into account when maximizing social welfare. Intuitively, erroneously computed counterfactual utility renders the coercion constraint a moving goal post, the trajectory of which the government has to consider when solving the coercion-constrained social welfare problem. Treating counterfactual utility as fixed, when it is, in fact, endogenous, leads either

[14]The same is true if the government itself is unable to correctly compute the full equilibrium at different taxes and consumers make the exact same mistakes as the government. Because we are working with a social planning model, the implicit assumption is that the government makes no mistake, but this assumption could be relaxed in a political economy framework where both candidates and voters may make mistakes, and which also could be similar in nature, e.g., because they rely on the same consulting firms to produce forecasts.

[15]With the exception of income levels, these parameters have been suppressed in the exposition, but they are implicit in the definition of preferences and the production possibilities frontier.

to a violation of the coercion constraint (if counterfactual utility increases), or to a non-binding coercion constraint and thus an unnecessary loss in welfare (if counterfactual utility decreases).

One example of such a situation is an emissions tax. It is likely that some people will likely under-estimate their expected utility under such a tax (if they do not account for revenue recycling or a decrease in negative externalities), whereas other people may over-estimate it (if they neglect the adverse consequences of price increases for all emissions-intensive goods and services). Once the tax is in place and prices adjust, both groups will (presumably) update their counterfactuals, although not to the extent that they agree on their assessment of a further tax reform. Another example could be a proposition for large-scale tax breaks, if some people over-estimate the net benefits (be they positive or negative) and others under-estimate them. In general, an adjustment in the counterfactual in response to a policy change can be expected to happen whenever people do not correctly anticipate the full consequences of a policy proposal.

2.2 Implications for welfare

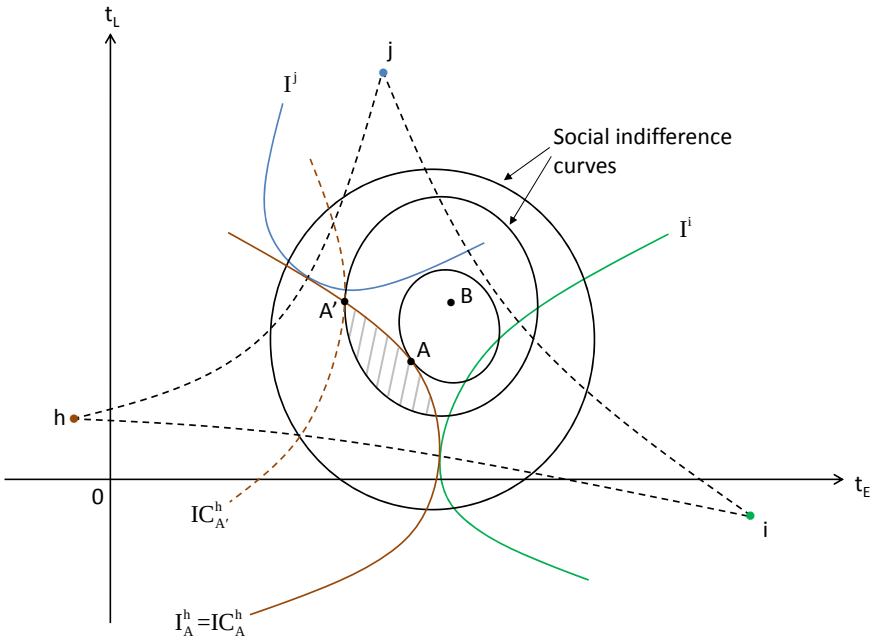
By construction, adding a binding coercion constraint cannot increase welfare as defined by the social welfare function.[16] Figure 1 illustrates, using a simplified setting where the government finances a public good by means of a labor tax t_L and a tax on an externality-generating good t_E. Every point in the figure reflects a specific policy combination. There are three consumer types $\{h, i, j\}$, who differ in their tastes and endowments, and their respective utility maxima are represented by the corners of the triangular shape.[17] The borders of this shape are the contract curves between any two types, which are defined by the loci of tangency of the indifference curves. We call the area enclosed by the contract curves the *Pareto-optimal policy space*, because any solution within this space is allocationally efficient and could be generated as a social welfare optimum based on a particular set of non-negative welfare weights. Conversely, any point outside the Pareto-optimal

[16] Recall that we do not capture potential losses from civil unrest or conflict in the social welfare function. If limiting coercion avoids severe social disruptions, it will increase welfare in a more general sense. As discussed in the introduction, this is an important reason to institute a coercion constraint in the first place.

[17] In order to span the entire policy space, we need at least three types. Limiting the example to three homogeneous types makes the analysis simpler, but it also reflects a normative concern that households should be treated equally by the government, as argued by Wallis (2014). Rather than gearing policy to individual consumers, the government only differentiates between broad classes of consumers in our setup. In our numerical model, these classes are workers and two types of capital owners.

policy space implies a negative welfare weight for at least one consumer type, which is inconsistent with a Paretian social welfare function.

Figure 1: Coercion-constrained policy outcome



Point B in Figure 1 represents the coercion-unconstrained welfare maximum (the "bliss point"), and the circular lines around it are social indifference curves. Suppose the coercion constraint is binding for type h, but not for types i and j (it will be binding for more than one group only as a special case). This type's indifference curves are circular lines around the utility maximum at the left corner of the policy space. An example of such an indifference curve is the line marked by I^h_A; examples of the other consumer types' indifference curves are I^j and I^i.

Under full rationality, type h's counterfactual utility coincides with his true utility maximum and is independent of actual policy, making the indifference curve also coincide with the corresponding iso-coercion curve.[18] The coercion-constrained policy outcome at point A is located at the tangency between the social indifference curve and type h's iso-coercion curve IC^h_A (which coincides with his indifference curve I^h_A) for a coercion level

[18] A completely correct counterfactual is a sufficient but not a necessary condition. A wrong counterfactual that is fixed, or a counterfactual that is wrong globally but correct locally satisfy the condition $\partial \tilde{V}^h / \partial t = 0$ as well.

of $V^h_{opt} - V^h_A = \bar{K}_h$.[19] The welfare loss from introducing the coercion constraint is the difference between welfare at the social optimum and the coercion-constrained outcome, W(B)-W(A). In order for the coercion constraint to make sense, this welfare loss has to be less than the societal gain from limiting coercion.

If consumers make mistakes when computing their counterfactual in a way that makes the latter endogenous with current policy, the iso-coercion curves will no longer be tangent to the iso-utility curves, because even a marginal move along an indifference curve changes the counterfactual and thus the level of coercion.[20] The coercion-constrained solution is again at the tangency between the social indifference curve and type h's relevant iso-coercion curve, for example $IC_{A'}$, leading to a solution at point A′. The resulting welfare loss in this example is W(B)-W(A′).

The endogeneity of the counterfactual leads to a welfare loss in itself. To see this, suppose we could somehow fix the counterfactual associated with A′ in consumer type h's mind, such that $\partial \tilde{V}^h / \partial t = 0$ and the iso-coercion curve coincides with the private indifference curve I_A. The resulting coercion-constrained policy outcome would be A as in the case of a correct counterfactual. The welfare loss due to the endogenous nature of the counterfactual is therefore W(A)-W(A′), whereas W(B)-W(A) can be interpreted as the unavoidable loss from introducing a coercion constraint. The move from policy A′ to policy A cannot take place, because it would make consumer h adjust his counterfactual utility such that the coercion constraint is exceeded. Correcting the error embedded in the calculation of the counterfactual allows society to reach point A, which provides a greater degree of welfare for the same level of coercion.

In general, welfare is increased by fixing the counterfactual whenever the indifference curve of the most-coerced consumer (i.e., the consumer type for whom the coercion constraint is binding) is not tangent to the iso-coercion curve at the equilibrium, which is the case if and only if $\partial \tilde{V}^h / \partial t \neq 0$.[21] Intuitively, the error embedded in the computation

[19]The introduction of coercion via an external constraint means that coercion does not count at all for groups that are coerced by less than $\bar{K}_h$. An alternative way to model a concern with coercion would be to include it into consumers' utility function. However, this would require a relative weighting of utility from private consumption vs. the disutility from being coerced for each group, and it is not clear on what basis this weighting would be chosen. Using an external coercion constraint also implies a relative weighting of welfare and coercion, but we would argue that it is easier to determine the maximum level of coercion that is acceptable in society than the rate at which each consumer trades off individual consumption against coercion.

[20]The slope of the indifference curve is $\frac{dt_L}{dt_E} = -\frac{\partial V / \partial t_E}{\partial V / \partial t_L}$, whereas the slope of the iso-coercion curve is given by $\frac{dt_L}{dt_E} = -\frac{\partial \tilde{V} / \partial t_E - \partial V / \partial t_E}{\partial \tilde{V} / \partial t_L - \partial V / \partial t_L}$. The two only coincide if $\frac{\partial \tilde{V}}{\partial t_L} = \frac{\partial \tilde{V}}{\partial t_E} = 0$.

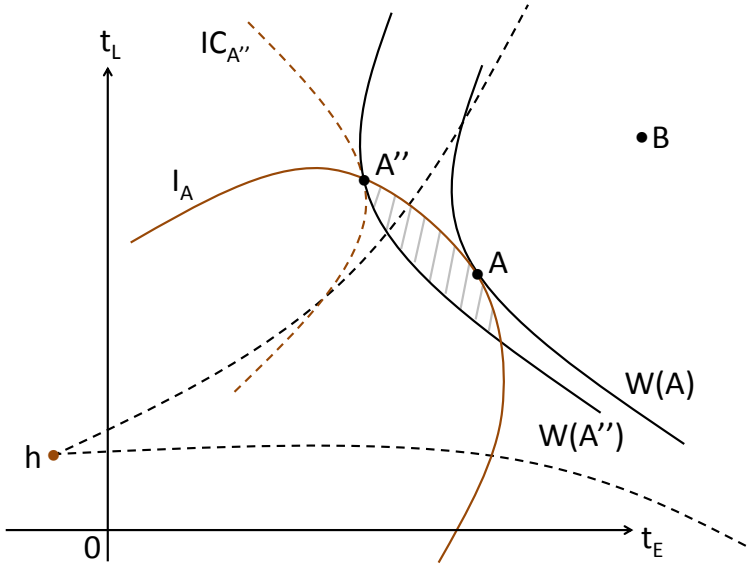
[21]To see this, note that pivoting the iso-coercion curve at point A′ in either direction allows for a policy solution that is at a higher social indifference curve than the one going through A′.

of the counterfactual utility can be interpreted as an additional constraint on the welfare maximization problem, and, like any constraint, it can only reduce welfare.

2.3 Efficiency

If consumer errors are large and the coercion constraint sufficiently tight, a biased counterfactual can lead to a policy equilibrium that is not only distributionally, but even allocationally inefficient. An example of such an outcome is point A″ in Figure 2, located at the tangency between the iso-coercion curve $IC_{A''}$ and the corresponding social indifference curve. Starting from point A″, moving inside the shaded area would not only increase social welfare and (weakly) type h's utility, but also allocational efficiency. The solution at A″ could not be replicated by a set of non-negative welfare weights, since all such solutions have to lie within the Pareto-optimal policy space.

Figure 2: Allocationally inefficient outcome



To understand the nature of the implicit negative welfare weight due to the introduction of a coercion constraint, we can interpret the additional terms in (8) as welfare weights that differ across policy dimensions. These weights can be positive or negative, depending on the coercion response to a change in a policy variable. Setting the LHS equal to zero and solving for α_h shows that consumer type h receives a negative welfare weight along policy dimension $t_i \in t$ if

$$\alpha_h < \kappa_h \cdot \left(\frac{\partial \tilde{V}^h / \partial t_i}{\partial V^h / \partial t_i} - 1 \right) \tag{15}$$

An inefficient result such as point A″ in Figure 2 implies that condition (15) applies to at least one policy dimension, and that the negativity of the dimension-specific welfare weight is sufficiently large to render the overall welfare weight negative. This condition can only occur if the coercion constraint for type h is binding, and if the marginal effect of t_i on counterfactual utility has the same sign and a greater magnitude than the effect on actual utility. Suppose that $\partial V^h/\partial t_i > 0$ at the equilibrium, because the utility gain via an increase in the public good more than offsets the utility loss from a higher price. If counterfactual utility, and thus coercion, increases by more than actual utility and α_h/κ_h is sufficiently low such that (15) holds, the government will consider this consumer type's preferences negatively when computing the optimal tax on X_i. The reason is that the increase in welfare due to an increase in the consumer's utility is more than offset by the increase in economic coercion.

The coercion-constrained outcome under an erroneous counterfactual can be interpreted as a type of paternalism, but one that refers to the consumers' counterfactual rather than actual utility functions. We refer to this situation as *counterfactual paternalism*: Not taking the changing nature of the counterfactual into account, consumer type h thinks he would prefer a move from A′ inside the shaded area in Figure 1, or from A″ inside the shaded area in Figure 2. The social planner does not allow such a move, even though it would increase this consumer's utility and overall welfare, because he knows that this would result in economic coercion beyond the limit $\bar{K}_h$, once the counterfactual adjusts.[22]

Our model is related to the "nudges" literature. The aim of nudges is to achieve a desirable outcome without coercion, e.g., by means of defaults; for a recent discussion, see Barton and Gruene-Yanoff (2015). In order for nudges to be welfare-improving, we have to assume that once people have been nudged into a certain choice, they end up being content and will not want to change their decision (even though they could). In other words, nudging people into a choice makes them realize that this is in fact the right choice for them, and this is qualitatively similar to our results: Only once a new policy equilibrium is reached, do consumers realize their actual level of coercion, and the social planner aims to avoid situations where people are coerced by more than they anticipate.

[22]Naturally, the concept of counterfactual paternalism implies that the government knows not only consumers' true optimal points, but also the nature of the counterfactual bias. If, on the other hand, the government expects a positive bias when the consumer in fact uses a negative one, a policy that aims to limit coercion would in fact exacerbate it.

2.4 Minimizing coercion

Unless consumers have homogeneous preferences and identical endowments, every policy choice will lead to some coercion since policy variables apply to everybody. A certain level of coercion is therefore unavoidable, even if the government were to single-mindedly pursue a policy of minimizing coercion. Naturally, economic coercion could be eliminated for some consumer group, but only at the price of significant coercion of others.

The minimum level of coercion that equally applies to all consumers can be found by minimizing $\bar{K}_h = \bar{K} \,\forall\, h$ subject to the production possibility frontier and the coercion constraint:[23]

$$\begin{aligned} \min_{t} \quad \bar{K} \qquad s.t. \quad & F(\cdot) \leq 0 \\ & \tilde{V}^h(\cdot) - V^h(\cdot) \leq \bar{K} \end{aligned} \tag{16}$$

Figure 3: Social welfare as a function of the constraint on coercion

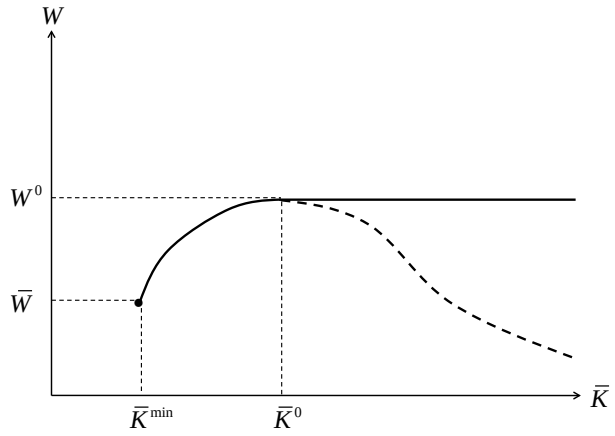


Figure 3 shows the relationship between the coercion constraint and social welfare. The slope of this figure is given by differentiating the government's Lagrangian w.r.t. the coercion constraint:

$$\frac{\partial W}{\partial \bar{K}} = \left.\frac{\partial \mathscr{L}}{\partial \bar{K}}\right|_{t^*} = \sum_{h=1}^{H} \kappa_h$$

[23] Even if all consumers are subject to the same numeric constraint, this could imply different levels of coercion if tastes and consumer errors are heterogeneous. Note that (16) looks similar to an equal (absolute) sacrifice problem. However, because the constraint is specified as an inequality, the absolute level of coercion will generally not be equalized across households. In contrast, application of the equal sacrifice concept in taxation leads to all consumers giving up the same (absolute, proportional or marginal) utility.

The minimum attainable level of coercion is given by $\bar{K}^{min}$, which is associated with a level of social welfare of $\bar{W}$. Coercion-unconstrained social welfare is maximized at the point $\left(\bar{K}^0, W^0\right)$, where all shadow prices κ_h are zero. Increasing $\bar{K}$ beyond this point has no effect on welfare in our model (solid line) as the constraint is no longer binding. If one were to impose a coercion constraint that holds with a strict equality as in Winer et al. (2014), rather than the weak inequality used in (6), increasing $\bar{K}$ beyond $\bar{K}^0$ would decrease social welfare as indicated by the dashed line. An outcome on this line would not be socially desirable, because lowering the level of coercion would both alleviate social tensions for the most coerced and increase overall welfare.

If, instead, the goal is to minimize coercion to one particular consumer (e.g., because this group is the most likely to engage in disruptive behavior or has the most political clout), the solution consists in choosing the tax combination that solves the following equation as a fixed point:

$$\mathbf{t}_{fp}^h = \text{argmax} \left\{ \tilde{\mathscr{L}}^h\left(\epsilon_h(\mathbf{t}_{fp}^h)\right) \right\} \tag{17}$$

Naturally, if consumers make no mistake when computing their counterfactual utility, the minimum level of coercion occurs at a consumer's true utility maximum. If, however, the counterfactual is endogenous to the policy from which it is computed, the consumer may not want to move away from a policy combination even if he would, in fact, be better off at a different equilibrium, because he mis-computes the utility associated with his true utility maximum. We will return to these "fixed points" in our numerical application.

3 Numerical illustration

To assess the magnitude of welfare changes from introducing a coercion constraint in the presence of consumer errors and to better understand the underlying mechanisms, we develop a numerical version of our model by choosing specific functional forms to represent preferences and technology. We simplify the theoretical model by restricting the number of policy instruments to a tax on labor and a tax on an intermediate good E that is associated with an aggregate environmental externality (for example, greenhouse-gas emissions affecting the climate), and to three consumer types. The model presented here is the non-calibrated version, which is easier for exposition purposes. For the actual

solution we transform everything into a calibrated share-form relative to a benchmark.

3.1 Specification of functional forms

We represent utility by a nested constant elasticity of substitution (CES) function. In the top nest, a private consumption composite C_h and leisure ℓ_h are combined according to a Cobb-Douglas functional relationship. This private aggregate then produces utility along with public provision G and environmental quality Q:

$$U^h = U^h(C_h, \ell_h, G, Q) = \left[\left(\theta_h^C + \theta_h^\ell\right)\left(C_h^{\delta_h}\ell_h^{1-\delta_h}\right)^{\rho_u} + \theta_h^G G^{\rho_u} + \theta_h^Q Q^{\rho_u}\right]^{1/\rho_u} \tag{18}$$

The parameter δ_h reflects the share in total income spent on private consumption goods within the Cobb-Douglas nest; the θ_h^k, $k = C, \ell, G, Q$, are the share parameters of the CES function with $\sum_k \theta_h^k = 1 \,\forall\, h$; and the exponent $\rho_u \equiv 1 - 1/\sigma_u$ reflects the curvature of the indifference curves, where σ_u is the elasticity of substitution between the private aggregate, public provision and environmental quality.

Consumers maximize their utility subject to the budget constraint

$$M_h = p_L L_h + I_h \leq C_h; \qquad L_h = \bar{L}_h - \ell_h \tag{19}$$

where L_h refers to consumer h's labor supply, p_L is the (uniform) net-of-tax wage and $\bar{L}_h$ represents the time endowment in efficiency units.[24] In addition to income from labor, consumers may also receive non-taxable income I_h.

Consumers' shadow value of time is given by ω_h. The comparative slackness condition

$$\omega_h \geq p_L; \quad L_h \geq 0; \quad (\omega_h - p_L) \cdot L_h = 0 \tag{20}$$

allows for the possibility of consumers not entering the labor market, if their valuation of time exceeds p_L. The consumer demand for private consumption and leisure resulting

[24] This allows us to use one wage rate that applies to all consumers, as consumer heterogeneity is captured by variation in $\bar{L}_h$.

from maximizing (18) subject to (19) is

$$C_h = \frac{\theta_h^C M_h}{p_C}$$
$$\ell_h = \frac{\theta_h^L M_h}{p_L} \qquad \text{if} \quad \omega_h \geq p_L; \qquad \ell_h = \bar{L}_h \ \text{ otherwise}$$

Production takes place in two stages. In the first stage, labor and I_h are used linearly to produce two intermediate goods:

$$\sum_h (L_h + I_h) = X + E \tag{21}$$

Good X is a clean composite, whereas E is associated with an externality that negatively affects environmental quality according to

$$Q(E) = E^{-\gamma} \qquad \gamma > 0$$

Due to the linear technology, producer prices for X and E are fixed, and we can choose quantities such that they are unity. In a second production stage, they are used to produce final output C and G according to a CES production function:

$$\begin{aligned} f(X,E) = Y &\equiv G + \sum_h C_h \\ &= \Phi\left[\theta_X X^{\rho_y} + (1-\theta_X)E^{\rho_y}\right]^{1/\rho_y}; \qquad \rho_y = 1 - 1/\sigma_y \end{aligned} \tag{22}$$

Here, $0 \leq \theta_X \leq 1$ is a share parameter, σ_y is the elasticity of substitution between X and E in the production of final output Y, and Φ is a constant that will take on the value of expenditure (including the valuation of public provision) in the benchmark to express utility in money-metric terms.

The government maximizes social welfare subject to a coercion constraint by choosing ad-valorem tax rates for labor and the dirty intermediate good, t_L and t_E, respectively, but it taxes neither the consumption composite nor the clean intermediate good X. Consumer prices for E and labor are given by $P_E = p_E(1 + t_E)$ and $P_L = p_L(1 + t_L)$, respectively. Tax revenue is used to fund the public good:

$$p_Y G = t_L p_L \sum_h L_h + t_E E \tag{23}$$

With CES technology, the price of the final output (equal to the unit cost of production, which applies to both C and G) is

$$p_Y = \frac{1}{\Phi}\left[\theta_X^{\sigma_y} + (1-\theta_X)^{\sigma_y}(1+t_E)^{1-\sigma_y}\right]^{1/1-\sigma_y} \tag{24}$$

Demand for the intermediate goods X and E can be derived by solving the cost minimization problem

$$\min_{X,E} \quad X + P_E \cdot E \qquad s.t. \quad f(X,E) \geq Y \tag{25}$$

with resulting demand functions of the form

$$X = \frac{Y}{\Phi} \cdot \left(\theta_X p_Y\right)^{\sigma_y} \tag{26}$$

$$E = \frac{Y}{\Phi} \cdot \left(\frac{(1-\theta_X)p_Y}{1+t_E}\right)^{\sigma_y} \tag{27}$$

Finally, market clearance and the government's budget constraint imply that

$$Y = \frac{\sum_h \omega_h L_h + I_h}{p_Y} + G \tag{28}$$

The key complexity in representing this model is writing down a system of equations which represents $\tilde{V}^h$, the welfare level which agent h believes he could receive if he were able to set taxes himself. That is, consumers compute their counterfactual level of utility by choosing the taxes on labor and the dirty good. We make the following assumption about consumers' counterfactual utility problem:

a. Consumers know their own utility function and budget constraint.

b. They recognize the public budget constraint and take into account the connection between tax revenue and the level of public provision.

c. However, when computing how policy instruments affect prices for goods and factors, aggregate levels of demand and supply, and thus their own welfare, they take a partial equilibrium view in the following sense:

   i. General equilibrium feedbacks on aggregate GDP are ignored when computing the effect of labor and energy taxes on total energy demand.

   ii. Agents treat the elasticities of the household-level and aggregate-level labor

supply as fixed, computed based on the benchmark, even though these elasticities will generally depend on the current equilibrium (because the shadow prices of labor and leisure change).

Our counterfactual model is a relatively small departure from full rationality. Larger errors would lead to much larger deviations between actual and consumer-computed counterfactual utilities. For example, if consumers did not respect the government's budget constraint, their counterfactual could be characterized by very low tax rates and high levels of public provision, or by high consumption and zero emissions, both of which are impossible given the available technology. We abstract from such "technology-inconsistent" counterfactuals.[25]

We define the following three consumer groups that differ with respect to their taxes and endowments:

- The "Brown Old" place a high valuation on public goods, but not on environmental quality. They have a high endowment of exogenous income I_h and a low labor endowment $\bar{L}_h$. This group's utility is maximized if the labor tax is high and the emissions tax is low.
- The "Green Young" have a high valuation for the environment and a relatively low valuation of other public goods. Their endowment is exclusively in terms of labor, and they can therefore be interpreted as the labor force. This group's utility is maximized if the labor tax is low and the emissions tax is high.
- The "Green Old" value both public goods and environmental quality highly, and have a high non-labor income and a low labor endowment. This group prefers high rates on both taxes.

3.2 Results

The top part of Table 1 shows the endowments and taste parameters for the three consumer groups. These parameters are chosen to generate significant heterogeneity with

[25]These types of counterfactuals were at the heart of our original model. However, their implementation is not straightforward, because either the counterfactuals are assumed to be fixed, in which case introducing a coercion constraint simply places more weight on the coercion-constrained group (see above), or they are assumed to change, but this requires a set of additional assumptions about how exactly they depend on the current equilibrium. In contrast, our partial-equilibrium representation of a general-equilibrium world generates a counterfactual that depends explicitly on the observed equilibrium via the elasticities of labor supply and the demand for the dirty good.

respect to endowments and tastes. This ensures that the different groups prefer very different policy outcomes, which in turn creates scope for economic coercion. Our parameter values are meant to capture the widespread disagreement with respect to taxation, public provision and environmental protection that we can observe in many countries today. We solve the model in calibrated share form and compute welfare, prices and quantities relative to the unconstrained welfare optimum. The full numerical model, including all parameter values and calibration steps, is available as a GAMS file in the online appendix.

Table 1: Model parameters and results

	Brown Old	Green Young	Green Old	Sum
V_G	3	0.25	3	
V_Q	0.25	2	2	
L_h	2	10	2	14
I_h	8	0	8	16
α_h	0.70	0.22	0.08	1
β_h	0.45	0.14	0.41	1
V^h at B	14.25	18.25	16.00	48.50
EV, in % of V(B)				
A' -> A	1.43	-3.59	0.00	-2.16
A' -> B	1.68	-2.43	-1.46	-2.21
A' -> FP	3.26	29.86	0.22	
A' -> GE Opt	6.31	29.98	1.13	
Welfare change, in % of V(B)				
A' -> A	1.00	-0.78	0.00	0.22
A' -> B	1.17	-0.53	-0.12	0.52

Note: V_G and V_Q represent the marginal valuation of the public good and environmental quality in the benchmark. Additional basic model parameters are $\sigma_u = \sigma_y = 0.5$, $\theta_X = 0.85$ and $\gamma = 2$. All other parameters are calibrated to the benchmark defined by $t_L = 0.5, t_E = 1$, subject to the chosen functional forms. The GAMS code for the full numerical model can be found in the online appendix.

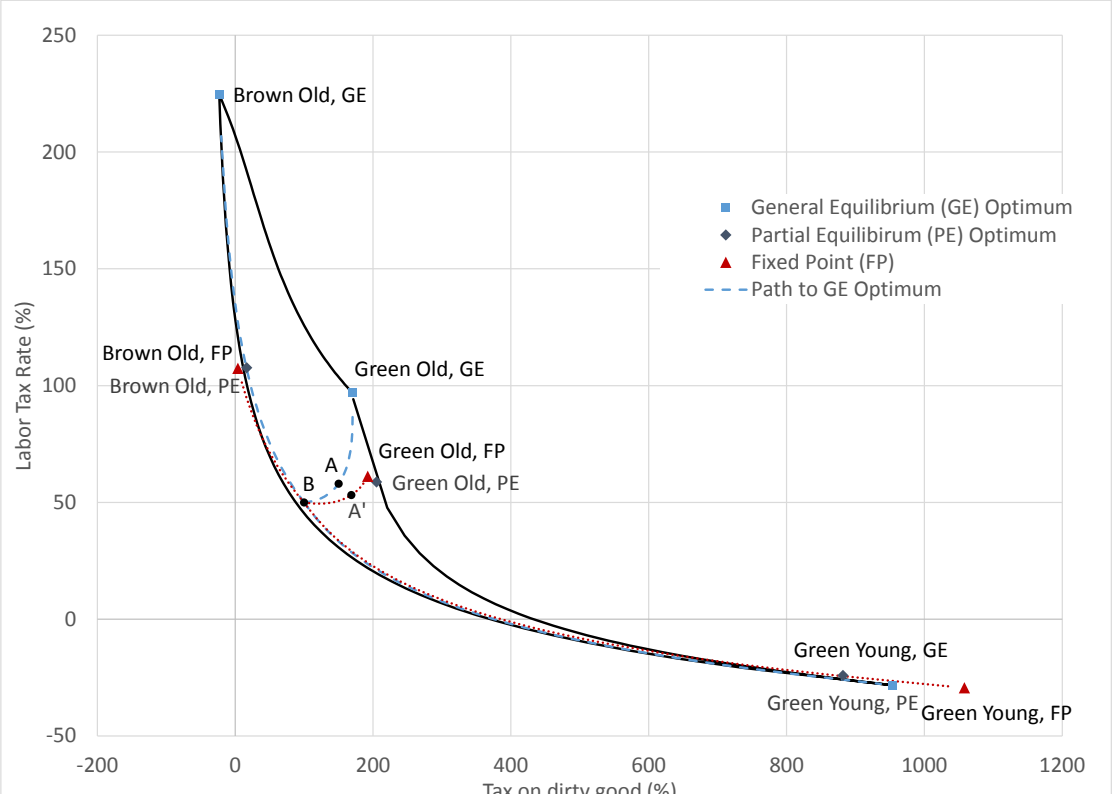
We choose as a benchmark the policy point defined by $t_L = 0.5$, $t_E = 1$ (i.e., a 50-percent tax on labor, and a 100-percent tax on the dirty good). This represents roughly the level of income and energy taxes in many European countries,[26] We use this benchmark as the coercion-unconstrained welfare maximum represented by point B in Figure 4, which is the numerical implementation of the (hand-drawn) Figure 1.

Starting from this benchmark, we compute the welfare weights α_h that need to be placed on the three consumer types in order for B to be the social optimum.[27] The paths

[26]For an overview of energy taxes around the world, see Parry et al. (2014).

[27]Given that the Green Young represent the workforce, one could argue for a higher welfare weight on this group. We chose the parametrization of the model mainly with a view to produce an intuitive figure in order to generate insights about limiting economic coercion, rather than to recreate a realistic distribution of the

Figure 4: Pareto-optimal policy space and policy equilibria



that lead from the benchmark to the GE optima are computed by gradually increasing the welfare weight on one of the agents, while holding the weights of the other two agents constant. Placing all the welfare weight on a single agent leads to the corners of the policy space, whereas placing all the weight on two agents while ignoring the utility of the third defines the contract curves between the corners. The partial equilibrium optima represent the consumer types' counterfactual equilibria as computed from B.

The introduction of a constraint that specifies that the Green Old's level of economic coercion must not exceed 50% of his coercion in the benchmark leads to the coercion-constrained equilibrium located at point A′, which is located at the tangency between the iso-coercion curve and the social indifference curve (not shown). Since this point is located within the policy space, it could also be achieved in the absence of a coercion constraint by using a different set of welfare weights; the weights that would give rise to a (coercion-unconstrained) solution at A' are given by β_h in Table 1.

The figure also shows the counterfactuals based on the partial equilibrium model (computed from point B), and the counterfactual "fixed points", which are the loci at which the policy equilibrium coincides with the respective consumer's counterfactual;

population. From a political economy point of view, one could argue that the workers may not have much lobbying power, and that therefore their preferences do not influence policy very much.

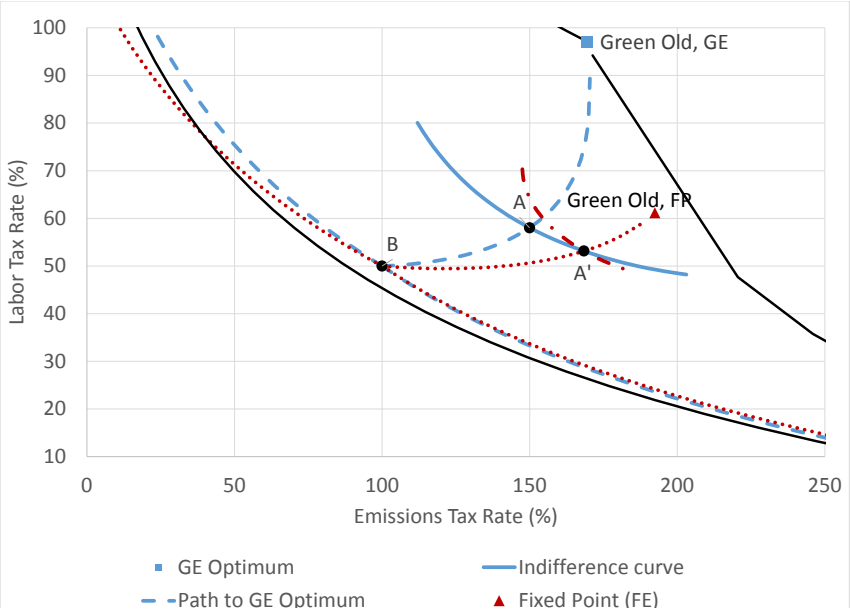
these fixed points are implicitly defined by (16). At these policy combinations (only), the respective consumer type would not perceive any economic coercion. Note that the fixed points do not coincide with the true GE optima due to the bias in the computation of the counterfactual.[28] Gradually reducing the maximum allowable level of coercion for one of the agents "pulls" the solution towards the fixed point; the approach paths computed by varying the coercion constraints between zero and the level at B for each agent are represented by the dotted lines in the figure. Point A′ is one example for placing a binding coercion constraint on the Green Old, but any point located on one of the three coercion approach paths represents a feasible coercion-constrained equilibrium, conditional on B being the unconstrained social optimum. This includes solutions outside the Pareto-optimal policy space as discussed in the theory section, for example by placing a very tight coercion constraint on the Brown Old.

Fixing the counterfactual utility of the Green Young at the level computed at A′ rotates the iso-coercion curve such that it coincides with the indifference curve, as shown in Figure 5. The coercion-constrained, though "error-corrected", equilibrium is located at point A, at the tangency between this indifference / iso-coercion curve and a some higher social indifference curve. The fact that point A lies on the optimal path between point B and the Green Old type's GE optimum illustrates the result that if consumers have correct (or equivalently, fixed) counterfactuals, placing a coercion constraint on a consumer group is equivalent to increasing its welfare weight relative to that of the other consumers. Last, if the social planner were to single-mindedly pursue a policy of coercion minimization based on a uniform coercion constraint for all agents, the equilibrium would be close to point A (the point is not shown for ease of exposition).

We derive welfare statements based on willingness to pay measures expressed in terms of the money-metric utility achieved at the benchmark. The middle panel of Table 1 shows the equivalent variation, as a percentage of benchmark income, associated with moving from A′ to another policy combination. Multiplying these equivalent variations by the welfare weights yields a measure for the change in social welfare associated with these policy changes, the values of which are shown in the bottom panel. The coercion constraint reduces social welfare by 0.52 percentage points, relative to the maximum achievable level

[28] Intuitively, consumers do not perceive a policy change towards their true optima as an improvement, because they over-estimate the costs and/or under-estimate the benefits of such a policy change by ignoring general-equilibrium effects.

Figure 5: Sensitivity of equilibrium to counterfactual bias



of welfare at point B and ignoring potential losses due to social tensions. Of this welfare loss, 0.22 percentage points are attributable to the erroneous nature of the counterfactual, whereas 0.30 percentage points are due to the unavoidable loss associated with imposing a coercion constraint itself. This implies that even the relatively mild departure from full rationality that we impose can significantly increase the welfare loss associated with limiting economic coercion.

Our numerical exercise is chosen with the aim of implementing a non-trivial coercion-constrained equilibrium that illustrates the main forces and concepts that we discuss in the theory section. We do not claim to reproduce an economic reality and therefore abstain from presenting alternative specifications, because there would be no way to judge which one is to be preferred. Using different parameter values or functional forms (including the shape of the coercion constraint) would naturally change the outcome of the numerical illustration.[29]

[29]Based on a suggestion made by an anonymous referee, we did, however, recompute our model after reducing the elasticity of labor supply to near zero (which simulates a lump-sum tax). Eliminating the distortive nature of the labor tax brings the PE equilibrium closer to the GE equilibrium and thus reduces the welfare loss implicit in limiting coercion and in consumers' bias. However, the qualitative nature of the equilibrium remains the same, and the resulting figure looks very similar to Figure 4. This means that it is not the distortive nature of the labor tax, but the disagreement about the size and distribution of the public budget that is at the root of the welfare losses inherent in limiting coercion under bounded rationality.

4 Conclusions

In this paper, we introduce an explicit concern with coercion in a Wicksellian sense, by departing from the assumption that people will simply accept the outcome of standard welfare maximization, regardless of the associated level of redistribution. We posit that people compare their actual utility, which is conditional on policy decisions that are exogenous to them, to a hypothetical situation where they are able to define policy themselves, and call the utility that people would derive in this hypothetical world their "counterfactual utility". The larger the difference between actual and counterfactual utility, the more people feel coerced by the economic reality they observe. Economic coercion is unavoidable if people differ with respect to their tastes or earning abilities, and it therefore cannot be a policy objective to eliminate economic coercion altogether. However, imposing a limit on the maximum acceptable level of economic coercion for particular consumer groups can be justified for reasons such as the protection of minorities, the wish to preempt socially damaging action from individuals who feel disenfranchised, or (in a political economy-context) to secure the favor of swing-voters. While governments typically employ negotiations to settle labor disputes and law enforcement to prevent and punish illegal activities, mitigating disenfranchisement ex-ante by limiting economic coercion may be worthwhile.

How much in "traditional" welfare (i.e., the welfare measured under the social planning paradigm that the solution to maximizing the social welfare function will be accepted by everybody) a society has to give up in exchange for limiting coercion depends crucially on whether people make mistakes when computing their counterfactual utilities. If they are completely rational and fully informed, their counterfactual utility becomes independent of actual policy decisions. Limiting coercion is then equivalent to increasing the welfare weight of particular consumers at the expense of others. The resulting equilibrium is different from the coercion-unconstrained case, but it is always allocationally efficient. The associated reduction in social welfare can be interpreted as the price society pays to reduce the risk of social strife below an acceptable level.

However, if consumers are not able to correctly compute their counterfactual utility, the latter becomes endogenous to current policy. Considering the complexity of modern economies and the multitude of interactions between different policies, prices and levels of demand and supply, an assumption of limited rationality seems natural. We show

that the welfare loss under incorrect counterfactuals exceeds the corresponding welfare loss under full rationality. In other words, consumers' errors give rise to a welfare loss beyond the unavoidable loss associated with limiting coercion. In our numerical model where we assume relatively minor consumer errors, the share of the welfare loss that is due to the error is about two-thirds. With a larger error, for example if consumers do not respect the government budget constraint or have a limited understanding about the available technology, this share would presumably increase. If consumer errors are sufficiently large and the coercion constraint sufficiently tight, the coercion-constrained policy outcome may even be allocationally inefficient in the sense that it could not be reproduced by any set of non-negative welfare weights (in the language used in this paper, the solution lies outside the Pareto-optimal policy space). We derive a necessary condition for this to occur.

Limiting economic coercion under bounded rationality can be interpreted as a particular form of paternalism, to which we refer as counterfactual paternalism. By this, we mean that there exist other policy choices that would increase the utility of the most-coerced consumer in addition to social welfare. However, changing policy in this direction would lead the consumer to adjust his counterfactual in a way that would lead to a violation of the coercion constraint and thus reduce his perceived welfare. The social planner does not have superior knowledge about the consumers' actual utility, but about their counterfactual utility, and he uses this knowledge to ensure that a predefined level of social discontent is not exceeded.

There are two important caveats to our results. First, the informational requirements for our model are substantial and exceed those from the standard framework, because the government needs to have information not only about consumers' utility, but also about the way in which they compute their counterfactual utility. Since the government has no magic wand of knowledge, it is likely to make an error when assessing consumers' bias in their counterfactual computations. Such an error would change the level of anticipated coercion (by either increasing or decreasing it, depending on the direction of the bias) and may thus itself become a source of coercion. However, in the context of public policy debates, the parties on the different sides of the argument often spell out precisely what type of consequences they expect from the policy change. To the extent that these statements can be taken at face value, they could provide valuable information about different

groups' counterfactuals.

Second, although our theory of social planning under a coercion constraint is general, our numerical results depend on the particular error that we assume consumers to make. To generalize the results and develop more intuition about the consequences of limiting coercion under bounded rationality, it would be worthwhile to examine various classes of errors and their effect on the coercion-constrained equilibrium, preferably in the context of a model that is calibrated to real economic data. This would allow for an assessment as to which types of common misperceptions the government should aim to reduce in order to limit economic coercion to some groups at the lowest social cost.

References

Arrow, Kenneth (1951). *Social Choice and Individual Values*. New Haven, CT Yale University Press.

Barton, Adrien and Till Gruene-Yanoff (2015). "From Libertarian Paternalism to Nudging-and Beyond." *Review of Philosophy and psychology* 6(3): 341–359.

Baumol, William J (2004). "Welfare Economics and the Theory of the State." in *The Encyclopedia of Public Choice*. Springer: 937–940.

Bernheim, B. Douglas and Antonio Rangel (2007). "Behavioral Public Economics: Welfare and Policy Analysis with Nonstandard Decision-Makers." in Peter Diamond and Hannu Vartiainen eds. *Behavioral Economics and its Applications*. Princeton University Press: 7–77.

Boadway, R. (2014). "Discussion: The Role of Coercion in Public Economic Theory." in J. Martinez-Vasquez and S. Winer eds. *Coercion and Social Welfare in Public Finance*. Cambridge University Press: 195–200.

Buchanan, James M. and Gordon Tullock (1962). *The Calculus of Consent. Logical Foundations of Constitutional Democracy*. University of Michigan Press.

Congleton, R. (2014). "Coercion, Taxation, and Voluntary Association." in J. Martinez-Vasquez and S. Winer eds. *Coercion and Social Welfare in Public Finance*. Cambridge University Press: 91–116.

Gamson, William A. (1961). "A Theory of Coalition Formation." *American Sociological Review* 26(3): 373–382.

Ledyard, J. (2014). "Non-Coercion, Efficiency, and Incentive Compatibility in Public Goods." in J. Martinez-Vasquez and S. Winer eds. *Coercion and Social Welfare in Public Finance*. Cambridge University Press: 143–159.

Lindahl, Eric (1919). "Just Taxation, a Positive Solution." in R. Musgrave and A. Peacock eds. *Classics in the Theory of Public Finance (1958)*. MacMillan: 168–176.

Parry, Ian, Dirk Heine, Eliza Lis and Sanjun Li (2014). *Getting Energy Prices Right: From Principle to Practice*. Washington, DC: International Monetary Fund.

Sehili, Saloua and Jorge Martinez-Vasquez (2014). "Lindahl Fiscal Incidence and the Measurement of Coercion." in J. Martinez-Vasquez and S. Winer eds. *Coercion and Social Welfare in Public Finance*. Cambridge University Press: 201–240.

Skaperdas, Stergios (1992). "Cooperation, Conflict, and Power in the Absence of Property Rights." *American Economic Review* 82(4): 720–739.

Skaperdas, S. (2014). "Proprietary Public Finance: On Its Emergence and Evolution Out of Anarchy." in *Coercion and Social Welfare in Public Finance*. Cambridge University Press: 60–81.

Wallis, J. (2014). "The Constitution of Coercion: Wicksell, Violence, and the Ordering of Society." in J. Martinez-Vasquez and S. Winer eds. *Coercion and Social Welfare in Public Finance*.: 29–59.

Wicksell, Knut (1896). "A New Principle of Just Taxation." in R. Musgrave and A. Peacock eds. *Classics in the Theory of Public Finance (1958)*. MacMillan: 72–118.

Winer, S., G. Tridimas and W. Hettich (2014). "Social Welfare and Coercion in Public Finance." in J. Martinez-Vasquez and S. Winer eds. *Coercion and Social Welfare in Public Finance*. Cambridge University Press: 160–194.