Journal of Economic Behavior and Organization 249 (2026) 107696

Contents lists available at ScienceDirect

Journal of Economic Behavior and Organization

journal homepage: www.elsevier.com/locate/jebo

Research paper

Universalization and the origins of fiscal capacity

Esteban Muñoz-Sobrado *

Department of Economics, Universitat Rovira i Virgili, Av. de la Universitat 1, Reus, 43204, Spain
ECO-SOS, Universitat Rovira i Virgili, Tarragona, Spain

ARTICLE INFO

Keywords:
State capacity
Universalization
Moral preferences

ABSTRACT

This paper proposes a model of tax compliance and fiscal capacity in which citizens partially internalize the consequences of concealment by imagining a world in which a share of the population acted similarly, linking their compliance decisions to the perceived value of public spending. A selfish elite chooses between public goods and private rents, taking compliance as given. In equilibrium, citizens' moral internalization expands the feasible tax base and induces elites to allocate resources toward provision rather than appropriation. When the value of public spending is uncertain, morality enables credible reform: high-value elites can signal their type through provision, prompting citizens to increase compliance and raising fiscal capacity within the same period. The analysis thus identifies a moral channel through which states may escape low-capacity traps even under weak enforcement.

1. Introduction

What sustains tax compliance and state capacity when enforcement is imperfect and institutions are fragile? Much of the economic literature emphasizes a coercive view of the state, where capacity depends on investments that extend its reach and allow it to enforce compliance (Besley and Persson, 2009, 2011). In contrast, a second view considers the emergence of a strong state as a collective arrangement in which citizens voluntarily comply with taxes in exchange for public goods and institutional legitimacy (Weingast, 1997, 2005; Binmore, 1994, 1998; Acemoglu, 2005). Within this view, economic models have microfounded compliance through intrinsic reciprocity, where reciprocal behavior between the elite and citizens is driven by citizens' preferences (Besley, 2020). Tax morale in such frameworks depends on perceptions of government behavior and can sustain either high- or low-capacity equilibria depending on expectations about the expenditure mix of the government.

This paper formalizes a different mechanism based on universalization reasoning, the idea that individuals evaluate their actions by asking, *What if everyone acted as I do?* We capture this logic using Homo Moralis preferences (Alger and Weibull, 2013, 2016), under which individuals evaluate outcomes as if a hypothetical share of the population acted the same way. The higher this share, the stronger is moral internalization. Full universalization is the limiting case. These preferences admit an evolutionary microfoundation but can be studied independently of it. In the context of taxation, this approach implies that individuals may comply even when enforcement is limited, because deviation carries an intrinsic moral cost as citizens internalize the universalized payoff implied by their compliance. The model shows that moral internalization can expand the fiscal frontier and foster the provision of public goods even under weak institutions. Moreover, when moral internalization is strong enough, credible improvements in fiscal behavior become feasible: self-interested elites find it optimal to allocate revenue to public goods rather than private rents, and citizens respond with higher compliance, thereby increasing fiscal capacity. By linking universalization reasoning to fiscal behavior, the

* Correspondence to: Department of Economics, Universitat Rovira i Virgili, Av. de la Universitat 1, Reus, 43204, Spain.
E-mail address: esteban.munoz@urv.cat.

https://doi.org/10.1016/j.jebo.2026.107696
Received 27 October 2025; Received in revised form 29 June 2026; Accepted 12 July 2026
Available online 20 July 2026
0167-2681/© 2026 The Author. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

analysis provides a microfoundation for self-enforcing taxation and offers a novel explanation of how states can move out of persistent low-capacity traps.

We build on the framework of Besley (2020), where an elite sets the tax rate and allocates revenue between public goods and self-serving transfers. The key departure lies in the behavioral foundation of compliance. In Besley's model, reciprocity enters through the composition of spending: civic-minded citizens comply more when the government allocates resources to public goods rather than rents, and civic-mindedness is a fixed trait. In our framework, compliance is driven by universalization: the moral return to reporting depends on the value of public spending α and institutional quality σ, not only on the spending mix. This structural dependence generates distinct predictions, particularly under asymmetric information, where universalization creates an informational channel absent under reciprocity.

We first explore these mechanisms in a static model with full information, establishing how morality shapes fiscal capacity and elite incentives, and then extend the framework to a dynamic setting in which the elite privately observes the value of public spending, so that its allocation choice serves as a signal and compliance responds to inferred fundamentals.

The static model delivers two main results. First, higher morality increases tax compliance, expanding the feasible revenue set and encouraging elites to allocate resources toward public goods rather than private rents. This mechanism supports high-tax, high-capacity equilibria. Second, moral preferences can sustain public provision even when the value of public goods is low, because sufficiently strong morality shifts the elite's incentive constraint in favor of provision. In this region, morality effectively disciplines elites into providing public goods that they would otherwise forgo, a role analogous to the cooperation incentives in Acemoglu (2005) but operating through moral internalization rather than repeated interaction.

To take the analysis beyond this static benchmark, we then study an infinite-horizon environment in which the value of public spending evolves over time and is privately observed by the elite. The elite's allocation choice becomes a signal of fiscal fundamentals: when citizens observe public spending rather than rents, they update their beliefs about the value of public goods. Under universalization, this update translates into higher compliance within the same period. This informational channel is absent under reciprocity, where compliance responds to the observed spending mix but not to inferred fundamentals. The dynamic structure thus produces a distinctive prediction, which we call *credible reform*: a self-enforcing equilibrium in which elite provision reveals a high-value state and the resulting belief update triggers an immediate expansion of fiscal capacity.

The dynamic analysis yields two main results. First, morality activates credible reform: when the value of public goods is only moderately high, self-interested elites would not provide them on material grounds alone, but strong enough moral internalization makes provision a credible signal of high productivity, prompting citizens to raise compliance within the same period. Second, morality amplifies the fiscal response to reform: higher morality magnifies the compliance jump triggered by credible provision, expanding the tax base immediately.

In our analysis, we distinguish between a weak-high state, in which morality is essential for credible reform, and a strong-high state, in which strong provision incentives make reform self-enforcing. Together, these findings suggest that moral societies have a structural advantage in escaping low-capacity equilibria once they experience positive shocks to the value of their common goods, an implication that does not follow directly from reciprocity-based models. We illustrate these mechanisms with two cases: the Nordic countries, where high moral internalization sustains compliance under moderate enforcement (Rothstein and Uslaner, 2005; Uslaner, 2008), and Botswana, where pre-existing civic norms enabled credible fiscal reform following the discovery of diamonds (Acemoglu et al., 2003; Robinson and Parsons, 2006).

Related literature. This paper relates to a broad literature that examines how cooperation and state authority can be sustained without perfect commitment. For instance, Kotlikoff et al. (1988) study self-enforcing intergenerational social contracts in an overlapping-generations framework, while Binmore (1998) models the endogenous formation of fairness norms through repeated interaction. Acemoglu (2005) develops a related framework in which a consensually strong state capable of taxation and rule enforcement emerges when the gains from cooperation outweigh temptations to expropriate. In these approaches, compliance and state capacity arise from strategic or institutional consistency through repeated play, reputation, or procedural rules that make cooperation incentive compatible. By contrast, the mechanism developed here does not rely on repetition or belief-based enforcement. It derives compliance from moral internalization, grounded in universalization reasoning, where individuals behave as if their own actions were to be generalized. This provides a belief-independent foundation for voluntary compliance and the emergence of fiscal capacity, complementing existing reciprocity-based explanations.

The paper also connects to work emphasizing the role of cultural traits such as trust and shared identity in shaping institutional performance (Algan and Cahuc, 2013; Collier, 2017), and more broadly to the literature on the positive effects of generalized trust on economic growth (Tabellini, 2010), and to research on the interaction between economic incentives and social norms in redistributive contexts (Lindbeck et al., 1999). These accounts typically model civic culture as generalized trust, in-group identification, or reciprocity norms. Our approach instead treats moral universalization as a preference that enters individual optimization directly, without requiring beliefs about others' compliance. Section 4.5 shows how the population share of this preference can itself be sustained endogenously through cultural transmission.

Finally, we contribute to a literature that applies Kantian reasoning to economic behavior, building on the early insight of Laffont (1975). Specifically, we draw on the Homo Moralis framework of Alger and Weibull (2013, 2016), in which individuals evaluate their actions as if universally adopted. This framework has been applied to settings such as team incentives (Sarkisian, 2017, 2021a,b), collective choice (Alger and Laslier, 2022; Alger et al., 2025), climate policy (Eichner and Pethig, 2021), Pigouvian taxation (Eichner and Pethig, 2022), monetary institutions (Norman, 2020), moral preferences in bargaining (Juan-Bartroli and Karagözoğlu, 2026),

injunctive norms (Juan-Bartroli, 2024), trade under information asymmetries (Rivero-Wildemauwe, 2025), norm emergence (Alger, 2025), and the interaction between moral concerns and social preferences (Juan-Bartroli and Rivero-Wildemauwe, 2025).[1]

The remainder of the paper is organized as follows. Section 2 introduces the baseline model and Homo Moralis preferences. Section 3 characterizes equilibrium under complete information, deriving the moral Laffer curve and the elite's optimal allocation. Section 4 extends the framework to asymmetric information, characterizes the signaling equilibria that produce the jump-start dynamic, and discusses robustness. Section 5 discusses the relation to reciprocity-based models, the behavioral foundations for moral universalization, and historical illustrations. Section 6 concludes.

2. Baseline model

We consider an economy with two types of agents: a ruling elite that sets fiscal policy, and a unit mass of identical tax-paying citizens. The elite acts as a collective decision-maker, while citizens take policy as given and decide how much to comply. All formal proofs are collected in the Mathematical Appendix.

Each citizen earns the same income $w > 0$. Citizens choose an income report $\tilde{w} \geq 0$ to the tax authority. Truthful reporting corresponds to $\tilde{w} = w$, while underreporting corresponds to $\tilde{w} < w$. Because citizens are ex-ante identical and atomistic, the equilibrium report is common across citizens (see the discussion following Lemma 2.1). We omit the citizen index throughout and write $\tilde{w}$ for this common report.

The elite sets fiscal policy to maximize its own payoff subject to institutional constraints, while compliance behavior is shaped by moral preferences introduced in the next subsection. The interaction takes the form of a two-stage game, formally defined in Section 3: the elite moves first, choosing fiscal policy. Citizens then observe the policy and choose income reports. In the dynamic extension (Section 4), the game is embedded in an infinite-horizon environment with asymmetric information.

The material payoff of each citizen depends on both public and private consumption and is assumed linear:

$$\pi(G, y) = \alpha \cdot G + y, \tag{1}$$

where G is a public good financed by tax revenue, y is private consumption, and $\alpha > 0$ captures the marginal value of the public good. For the baseline analysis we assume α is common to citizens and the elite. Appendix B extends the analysis to the unaligned case, in which the elite and citizens have distinct valuations α_E and α_C, respectively.

2.1. Policy and institutions

The elite sets fiscal policy, which consists of four elements: the income tax rate t, public good provision G, transfers to the elite B, and transfers to citizens b. Institutional strength is captured by a parameter $\sigma \in (0, 1)$, which governs the extent to which elite appropriation must be matched by transfers to citizens. For every unit appropriated by the elite, institutions require σ units to be transferred to citizens, so that $b = \sigma B$. The elite's effective share of residual revenue is $\theta(\sigma) = 1/(1+\sigma)$, strictly decreasing in σ, with $\theta(0) = 1$ and $\theta(1) = 1/2$.

This specification follows Besley and Persson (2011). The parameter σ summarizes the accountability institutions that constrain elite appropriation. At $\sigma \to 0$ diverted revenue flows entirely to the elite. At $\sigma \to 1$ each unit diverted is matched one-for-one by a citizen transfer, leaving the elite with half. The composite $\sigma\,\theta(\sigma) = \sigma/(1+\sigma)$, the fraction of diverted revenue recaptured by citizens as transfers, will recur throughout the analysis.

For analytical convenience, we parameterize the allocation of tax revenue by a share $g \in [0, 1]$, so that $G = gT$ and $B = \theta(\sigma)(1-g)T$, where T denotes aggregate tax revenue. The elite's problem then reduces to a choice over (t, g).

2.2. Compliance under universalization

Citizens choose a reported income $\tilde{w} \geq 0$ to the tax authority. Post-tax, pre-transfer income is

$$z(\tilde{w}) = w - t\,\tilde{w} - c\,C(\tilde{w} - w), \tag{2}$$

where $c > 0$ measures enforcement intensity and $C : \mathbb{R} \to \mathbb{R}_+$ is a convex misreporting-cost function of the deviation $d = \tilde{w} - w$, with $C(0) = 0$, $C'(0) = 0$, and $C''(d) > 0$.[2] Final private consumption is

$$y(\tilde{w}) = b + z(\tilde{w}). \tag{3}$$

[1] A companion paper, Muñoz-Sobrado (2022), applies Homo Moralis preferences to optimal income taxation. The present paper focuses instead on the emergence and persistence of fiscal capacity. López-Cantor (2025) analyzes the interplay between individual morality, public good provision, and redistribution in a static political economy setting, complementing the present analysis by examining the distributional consequences of Kantian moral preferences through voting equilibria.

[2] This reduced-form cost admits two interpretations. Under a *material* reading, $cC(\tilde{w}-w)$ is the expected penalty from detection, and c indexes the government's detection effort, in parallel with the strategic investments in coercive capacity studied by Besley and Persson (2009, 2011), Besley (2020). Under a *lying-cost* reading, it is an intrinsic psychic cost of dishonesty, and c scales the citizen's aversion to untruthful reporting. Because the term enters the citizen's actual and universalized net-income expressions symmetrically, the two readings yield identical first-order conditions and equilibrium objects. We adopt the material reading as the default.

Throughout the analysis, we adopt the quadratic specification $C(d) = \frac{1}{2}d^2$, which delivers closed-form solutions while preserving the key convexity properties.[3]

Universalization benchmark. If all citizens report the same $\tilde{w}$, per-capita objects are

$$\begin{aligned} T^{\mathcal{M}}(\tilde{w}) &= t\,\tilde{w} && \text{(Tax revenue)} \\ G^{\mathcal{M}}(\tilde{w}) &= g\,T^{\mathcal{M}}(\tilde{w}) && \text{(Public good)} \\ b^{\mathcal{M}}(\tilde{w}) &= \sigma\,\theta(\sigma)\,(1-g)\,T^{\mathcal{M}}(\tilde{w}) && \text{(Transfers to citizens)} \\ z^{\mathcal{M}}(\tilde{w}) &= w - t\,\tilde{w} - c\,C(\tilde{w}-w) && \text{(Net income).} \end{aligned}$$

Definition 2.1 (*Homo Moralis Utility*)**.** Each citizen is characterized by a degree of morality $\kappa \in [0,1]$.[4] Given a report $\tilde{w}$, the citizen's utility is

$$U^{(\kappa)}(\tilde{w}) = (1-\kappa)\,\pi(G,\, b + z(\tilde{w})) + \kappa\,\pi\left(G^{\mathcal{M}}(\tilde{w}),\, b^{\mathcal{M}}(\tilde{w}) + z^{\mathcal{M}}(\tilde{w})\right), \tag{4}$$

where $\pi(G,y) = \alpha G + y$ is the material payoff function.

The Homo Moralis utility is a convex combination of two payoff evaluations. The citizen attaches weight $1-\kappa$ to a selfish evaluation, in which policy is taken as given, and weight κ to a universalized evaluation, in which she assesses the public good provision and transfers that would obtain if everyone submitted the same report.[5]

Given the macroeconomic constraints and the linear tax structure, the fiscal variables (public good provision G, transfers to citizens b, and elite rents B) can be written as functions of the tax rate t, the public good share g, and citizens' reports. For any given policy parameters (t,g), each citizen chooses a report $\tilde{w} \geq 0$ to solve

$$\begin{aligned} \tilde{w}^*(\kappa; g,\alpha,t,c,\sigma) = \arg\max_{\tilde{w}\geq 0}\ & (1-\kappa)\,\pi(G,\, b + z(\tilde{w})) \\ & + \kappa\,\pi\left(G^{\mathcal{M}}(\tilde{w}),\, b^{\mathcal{M}}(\tilde{w}) + z^{\mathcal{M}}(\tilde{w})\right), \end{aligned}$$

where $\pi(G,y) = \alpha G + y$ is the material payoff function and

$$z(\tilde{w}) = w - t\,\tilde{w} - c\,C(\tilde{w}-w) \tag{5}$$

denotes post-tax, pre-transfer income.

The following lemma characterizes the citizen's optimal report under the quadratic specification.

Lemma 2.1 (*Optimal Report*)**.** *Under the quadratic specification $C(d) = \frac{1}{2}d^2$, any interior solution to the citizen's problem satisfies*

$$\tilde{w}^*(\kappa; g,\alpha,t,c,\sigma) = w + \frac{t}{c}\left[\kappa\,\varphi(g,\alpha,\sigma) - 1\right], \tag{6}$$

where

$$\varphi(g,\alpha,\sigma) := g\,\alpha + (1-g)\,\sigma\,\theta(\sigma)$$

is the effective moral return to reporting, a weighted average of the marginal value of public goods and transfers under universalized behavior.

Since each citizen is atomistic and ex ante identical, and π is linear in G and b, Lemma 2.1 yields a unique best reply common to all citizens. As a result, the Stage 2 equilibrium is symmetric.[6]

Compliance behavior is governed by the product $\kappa\,\varphi(g,\alpha,\sigma)$. The effective moral return φ aggregates the value of the public good, the allocation share to provision, and the degree of institutional matching, so that moral agents respond endogenously to the government's expenditure mix. If $\kappa\,\varphi(g,\alpha,\sigma) < 1$, citizens under-report. If $\kappa\,\varphi(g,\alpha,\sigma) = 1$, they report truthfully. If $\kappa\,\varphi(g,\alpha,\sigma) > 1$, they over-comply. The benchmark $\kappa = 0$ reproduces the purely selfish case in which citizens always under-report, with the extent of concealment determined only by enforcement c and the tax rate t.

In what follows we normalize income to $w = 1$ without loss of generality. We discuss this point when deriving the elite's allocation rule (Proposition 3.1).

Assumption 1. The degree of morality satisfies $\kappa\,\alpha < 1 - c/2$.

[3] This specification penalizes deviations from truthful reporting symmetrically. Since Assumption 1 below ensures that the optimal report never exceeds true income, the cost of over-reporting is never incurred in equilibrium and the symmetric form is without loss.

[4] We follow the convention of Alger and Weibull (2013): $\kappa = 0$ is the Homo Economicus benchmark and $\kappa = 1$ is the pure-Kantian limit. The regularity bound $\kappa\alpha < 1 - c/2$ in Assumption 1 below ensures that the Laffer-maximizing tax rate is interior and revenue remains finite.

[5] This counterfactual ties individual compliance decisions to the macroeconomic constraints of the economy, in the sense of Laffont (1975).

[6] Since the optimal report in Lemma 2.1 is linear in κ, aggregate compliance in a heterogeneous population depends only on the mean degree of morality. The representative-agent specification is therefore without loss of generality for aggregate fiscal capacity. Appendix D develops the two-type case with endogenous shares.

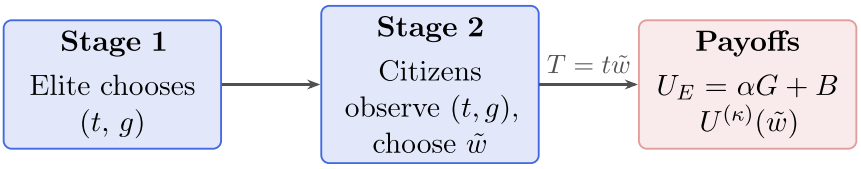


Fig. 1. Timing of the static fiscal game $\Gamma(\alpha, \kappa, c, \sigma)$. The elite moves first, choosing fiscal policy (t, g). Citizens then observe policy and choose income reports $\tilde{w}$. Payoffs are realized after revenue $T = t\tilde{w}$ is generated. All parameters $(\alpha, \kappa, c, \sigma)$ are common knowledge, and the game is solved by backward induction.

This restriction does two things. First, it rules out over-compliance: whenever $\kappa\alpha \geq 1$, the universalized return to contribution under $g = 1$ weakly dominates unity, so reported income weakly exceeds true income and feasible revenue is unbounded. Second, and strictly stronger, it ensures that the Laffer problem admits an interior solution: the revenue-maximizing tax rate $\hat{t}(g, \kappa, c, \sigma)$ defined in (8) satisfies $\hat{t} < 1$ for every feasible g, so the mechanism operates in the regime where citizens respond to tax-rate changes through compliance adjustments rather than corner outcomes. All qualitative results are continuous in κ and hold for any κ bounded away from the boundary $\kappa\alpha = 1 - c/2$.

3. The static framework

This section characterizes equilibrium under complete information about the value of public goods. We first define the game formally, then solve for citizens' compliance and the resulting fiscal capacity, and finally characterize the elite's optimal allocation.

3.1. Game form and equilibrium concept

Definition 3.1 (*Static Fiscal Game*)**.** The *static fiscal game* $\Gamma(\alpha, \kappa, c, \sigma)$ is defined by:

1. **Players.** A ruling elite E, formalized as a single decision-maker, and a unit mass of identical citizens $i \in [0, 1]$.
2. **Timing.**

Stage 1. The elite publicly chooses a fiscal policy $(t, g) \in [0, 1]^2$, where t is the tax rate and g is the share of revenue allocated to the public good. Transfers B and b are determined residually by the budget constraint and the institutional parameter σ.

Stage 2. Each citizen i observes (t, g), and all citizens simultaneously choose reports $\tilde{w}_i \geq 0$.

3. **Payoffs.**

- Each citizen's payoff is the Homo Moralis utility $U^{(\kappa)}(\tilde{w}_i)$ defined in Eq. (4).
- The elite's payoff is $U_E = \alpha G + B$, where $G = gT$ and $B = \theta(\sigma)(1 - g)T$ are determined by aggregate revenue $T = \int_0^1 t\,\tilde{w}_i\,di$.

4. **Information.** All parameters $(\alpha, \kappa, c, \sigma)$ are common knowledge.

As noted after Lemma 2.1, each citizen's best reply is independent of other citizens' reports, so the Stage 2 equilibrium is symmetric and coincides with the individual optimum characterized there. The solution concept is *Subgame Perfect Equilibrium* (SPE): in Stage 2, each citizen plays the unique best response $\tilde{w}^*(\kappa; g, \alpha, t, c, \sigma)$ from Lemma 2.1. In Stage 1, the elite anticipates this response and chooses (t^*, g^*) to maximize U_E. Fig. 1 summarizes the timing.

3.2. Fiscal capacity and the laffer curve

Fiscal capacity is defined as the maximum tax revenue the government can raise, given citizens' degree of morality κ and the enforcement parameter c. Per-capita tax revenue, given a tax rate t and the reporting choice $\tilde{w}(\kappa; g, \alpha, t, c, \sigma)$, is

$$\begin{aligned} T(t, g, \kappa, c, \sigma) &= t\,\tilde{w}(\kappa; g, \alpha, t, c, \sigma) \\ &= \frac{t}{c}\Big[c - t\big(1 - \kappa\,\varphi(g, \alpha, \sigma)\big)\Big]. \end{aligned} \tag{7}$$

This expression yields a Laffer-type curve: tax revenue is hump-shaped in the tax rate, rising approximately linearly at low t and eventually declining as reduced compliance offsets the mechanical rate increase.

The revenue-maximizing tax rate solves $\max_t T(t, g, \kappa, c, \sigma)$, yielding

$$\hat{t}(g, \kappa, c, \sigma) = \frac{c/2}{1 - \kappa\,\varphi(g, \alpha, \sigma)}. \tag{8}$$

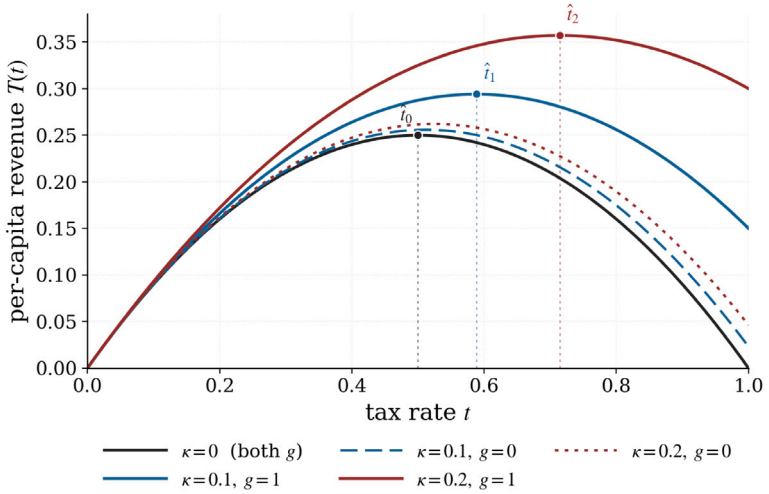


Fig. 2. Moral Laffer curves. Per-capita tax revenue $T(t)$ as a function of the tax rate for three morality levels $\kappa \in \{0, 0.1, 0.2\}$ and two allocation policies $g \in \{0, 1\}$, evaluated at $\alpha = 1.5$, $\sigma = 0.3$, and $w = c = 1$. At $\kappa = 0$ the two allocation policies yield identical revenue, shown as a single solid black line. For $\kappa > 0$, solid lines denote full provision ($g = 1$) and dashed lines denote full rents ($g = 0$). Filled dots mark the revenue-maximizing tax rate $\hat{t}$ on each $g = 1$ curve, with vertical dashed lines indicating its position on the tax-rate axis.

Substituting into (7) gives the peak of the Laffer curve:

$$T(\hat{t}, g, \kappa, c, \sigma) = \frac{c}{4\,[1 - \kappa\,\varphi(g, \alpha, \sigma)]}. \tag{9}$$

Fig. 2 plots these curves for several values of κ and both allocation policies $g \in \{0, 1\}$.

Three features of Fig. 2 bear on the mechanism developed below. First, at $\kappa = 0$ compliance depends only on the tax rate and enforcement, so the allocation of revenue is irrelevant: the Laffer curve is fixed. Second, as κ rises the whole curve shifts up and the revenue-maximizing rate $\hat{t}$ shifts to the right, jointly enlarging the height and location of fiscal capacity. Third, and most important for what follows, the gap between the $g = 1$ (solid) and $g = 0$ (dashed) curves widens with κ: this is the *moral premium for provision*. In the figure, the $g = 0$ curves sit close to the $\kappa = 0$ baseline because at these parameters the moral return under rents ($\kappa\sigma\theta(\sigma)$) is several times smaller than the moral return under provision ($\kappa\alpha$). Because $\varphi(1, \alpha, \sigma) = \alpha$ substantially exceeds $\varphi(0, \alpha, \sigma) = \sigma\theta(\sigma)$ when institutions are weak, morality couples fiscal capacity to the allocation decision. This coupling is what generates the elite discipline characterized in Proposition 3.1: the tax base the elite can extract is not a primitive of the economy but a function of how it chooses to use revenue.

3.3. The elite's problem

The elite's utility is given by $U_E = \alpha G + B$, where G denotes public good provision and B rents. The elite chooses (t, g) to maximize this utility. Conveniently, the problem can be re-expressed as a choice over $g \in [0, 1]$, the share of tax revenue allocated to the public good.

For each g, the tax rate $\hat{t}(g, \kappa, c, \sigma)$ is chosen to maximize revenue, and the resulting allocations satisfy

$$G(g) = g \cdot \hat{T}(g), \qquad B(g) = \theta(\sigma)\,(1 - g)\,\hat{T}(g), \tag{10}$$

where

$$\hat{T}(g) := T(\hat{t}(g, \kappa, c, \sigma), g, \kappa, c, \sigma) \tag{11}$$

denotes per-capita revenue under the Laffer-maximizing tax rate at allocation share g.

Using the constraints in Eq. (10), the elite's problem reduces to $\max_{g \in [0,1]} U_E(\alpha; g)$, where

$$U_E(\alpha; g) = \hat{T}(g)\,[\alpha g + \theta(\sigma)(1 - g)]. \tag{12}$$

Because the tax rate is optimized for each g, the problem reduces to a one-dimensional maximization over the allocation share g. The elite's objective $U_E(\alpha; g)$ is the product of two functions that are affine in g, so it is single-peaked and the optimum is always at a corner. The following result characterizes which corner obtains.[7]

[7] Having defined the elite's objective, we elaborate on the normalization $w = 1$. For general w, the misreporting cost $C(d) = d^2/2$ depends on the deviation $d = \tilde{w} - w$ alone, so equation (7) reads $T = (t/c)\,[wc - t(1 - \kappa\varphi)]$, the revenue-maximizing rate is $\hat{t} = wc/[2(1 - \kappa\varphi)]$, and the peak of the Laffer curve (9) is $\hat{T}(g) = w^2 c/[4(1 - \kappa\varphi(g, \alpha, \sigma))]$. Since revenue enters the elite's objective (12) multiplicatively, general w scales that objective by the positive constant w^2. The normalization thus does not affect the elite's optimal strategy or the thresholds characterized below. A more general treatment would require Assumption 1 to read $\kappa\alpha < 1 - wc/2$.

Proposition 3.1 (*Elite's Allocation Under Alignment*). *The elite's optimal allocation* $g^* \in \{0, 1\}$ *satisfies:*

1. ***Strong provision state.*** *If* $\alpha > \theta(\sigma)$*, provision is always optimal, i.e.* $g^* = 1$ *for all feasible* κ*.*
2. ***Weak provision state.*** *If* $\sigma\,\theta(\sigma) < \alpha \leq \theta(\sigma)$*, there exists a unique morality threshold*

$$\bar{\kappa}(\alpha, \sigma) = \frac{\theta(\sigma) - \alpha}{\alpha\,\theta(\sigma)\,(1 - \sigma)} \in \big(0,\ 1/\alpha\big), \tag{13}$$

such that

$$g^* = \begin{cases} 0, & \text{if } \kappa < \bar{\kappa}(\alpha, \sigma), \\ 1, & \text{if } \kappa \geq \bar{\kappa}(\alpha, \sigma). \end{cases} \tag{14}$$

3. ***Transfer state.*** *If* $\alpha \leq \sigma\,\theta(\sigma)$*, the elite always prefers rents, i.e.* $g^* = 0$ *for all feasible* κ*.*

Moreover, the threshold $\bar{\kappa}(\alpha, \sigma)$ *satisfies:*

$$\frac{\partial \bar{\kappa}}{\partial \alpha} < 0, \qquad \frac{\partial \bar{\kappa}}{\partial \sigma} \begin{cases} < 0 & \text{if } \alpha > \frac{1}{2}, \\ = 0 & \text{if } \alpha = \frac{1}{2}, \\ > 0 & \text{if } \alpha < \frac{1}{2}. \end{cases} \tag{15}$$

Hence higher α *always reduces the morality required to induce provision, and stronger institutions reduce (resp. increase) the required morality when* $\alpha > \frac{1}{2}$ *(resp.* $\alpha < \frac{1}{2}$*).*

The elite's allocation determines the equilibrium tax base, which responds to both the spending mix and citizens' morality. The three regions in Proposition 3.1 admit a clear economic interpretation. In the *strong provision state* ($\alpha > \theta(\sigma)$), the elite's valuation of the public good exceeds the residual rents from diversion, and provision is chosen unconditionally. Morality still affects the magnitude of the tax base but is inessential for the allocation itself. The *transfer state* ($\alpha \leq \sigma\theta(\sigma)$) is the mirror case: the public good is valued so weakly that no admissible morality can make provision optimal, and the elite diverts throughout. The paper's mechanism operates in the intermediate *weak provision state*, where the elite's private ranking favors rents but lies close enough to indifference that morality can flip the allocation. Once citizens' internalization κ crosses $\bar{\kappa}(\alpha, \sigma)$, universalized reasoning expands the tax base under $g = 1$ enough to overturn the elite's preference for rents, and provision becomes optimal.

Corollary 3.1 (*Taxes Raised in Equilibrium*). *At the elite's optimum, the equilibrium tax base is given by*

$$T_1(\kappa, \sigma) = \frac{c}{4\,(1 - \kappa\,\alpha)} \qquad \text{if } g^* = 1, \tag{16}$$

$$T_0(\kappa, \sigma) = \frac{c}{4\,(1 - \kappa\,\sigma\,\theta(\sigma))} \qquad \text{if } g^* = 0. \tag{17}$$

Denote by $T^*(\kappa, \sigma)$ *the tax base at the elite's optimal allocation* g^**.*

Two features of the threshold $\bar{\kappa}(\alpha, \sigma)$ deserve emphasis. First, $\partial\bar{\kappa}/\partial\alpha < 0$: morality and fundamentals are *substitutes* in inducing provision. Societies with a low value of public spending can compensate through strong internalized norms, and vice versa. This substitutability is the basis of the moral channel identified in the introduction: the ability of weak states to sustain fiscal capacity through citizen universalization rather than coercive enforcement alone. Second, $\partial\bar{\kappa}/\partial\sigma$ changes sign at $\alpha = 1/2$: morality and institutions are *complements* when the public good is sufficiently valuable ($\alpha > 1/2$) and *substitutes* otherwise. In the complementary regime, better institutions *reduce* the morality required to induce provision: coercive capacity and moral capacity work in tandem to discipline the elite. In the substitutive regime, stronger institutions *raise* the threshold: the two channels compete for doing the same work, and one partially crowds out the other.[8]

Beyond the allocation margin, the equilibrium tax base $T^*(\kappa, \sigma)$ is monotone in κ: moral agents comply more, expanding capacity regardless of the allocation. The expansion is greatest under provision, since the denominator $1 - \kappa\alpha$ in the $g^* = 1$ regime falls more steeply in κ than $1 - \kappa\sigma\theta(\sigma)$ does in the $g^* = 0$ regime. Stronger morality thus operates on two margins simultaneously: it amplifies the fiscal returns to provision and shifts the elite's optimal allocation toward it, disciplining elites into providing public goods even when $\alpha < \theta(\sigma)$.

4. Dynamics and asymmetric information

The static model takes the value of the public good as given. In practice, fiscal fundamentals shift over time, and the elite, through privileged access to administrative data, fiscal projections, and policy expertise, may be better informed than citizens about the current state. This section extends the framework to an infinite-horizon environment in which the elite privately observes the value of public spending each period, so that its allocation choice becomes a potential signal of fiscal fundamentals.

The central message is that moral universalization enables credible fiscal reform under asymmetric information, even when institutions are weak. High-value elites signal their type through provision, citizens infer the improvement in fundamentals and

[8] This regime distinction is only relevant where the threshold is interior. Under Assumption 1, that requires α to exceed $\alpha_{\min}(c, \sigma) = \theta(\sigma)\,[1 - (1 - c/2)(1 - \sigma)]$. Notably, $\alpha_{\min}$ equals $1/2$ when $c = 1$, for every σ, exactly the value at which the complementary and substitutive regimes switch.

raise compliance, and the tax base expands within the same period. Lemma 4.1 shows that equilibrium behavior in this signaling environment admits a tractable stage-game characterization, and Propositions 4.1–4.2 identify the morality thresholds that sustain credible reform together with the resulting fiscal multiplier.

4.1. Environment and timing

The value of the public good follows a two-state Markov chain privately observed by the elite at the start of each period.[9] As in the static model, Assumption 1 restricts morality to $\kappa\,\alpha^H < 1 - c/2$.

Definition 4.1 (*Dynamic Fiscal Signaling Game*)**.** The *dynamic fiscal signaling game* extends Γ to an infinite-horizon setting with asymmetric information and time-varying fundamentals. Primitives, technology, and action spaces are as in Definition 3.1. Time is discrete, $n = 0, 1, 2, \ldots$. Each period proceeds in three stages.

Stage 0. Nature draws the value of the public good $\alpha_n \in \{\alpha^L, \alpha^H\}$ from a Markov chain with transition probabilities $q^H \equiv \Pr(\alpha_{n+1} = \alpha^H \mid \alpha_n = \alpha^H)$ and $q^L \equiv \Pr(\alpha_{n+1} = \alpha^H \mid \alpha_n = \alpha^L)$, with $q^H \geq q^L$. The chain is initialized at its stationary distribution, so the marginal probability is

$$\Pr(\alpha_n = \alpha^H) \;=\; \rho \;\equiv\; \frac{q^L}{1 - q^H + q^L} \;\in\; (0,1)$$

each period. At each n the elite privately observes α_n.[10]

Stage 1. The elite, knowing α_n, publicly chooses a fiscal policy $(t_n, g_n) \in [0,1]^2$.

Stage 2. Citizens observe (t_n, g_n), form a posterior p_n over α_n by Bayes' rule (on path) and the D1 refinement (off path), and choose reports $\tilde{w}_n$ to maximize the Homo Moralis utility $U^{(\kappa)}$ in Eq. (4), with universalization applied to the stage-game report rather than to a history-contingent strategy.[11]

Strategies. Strategies are Markov. The elite's strategy is a mapping $\gamma_E : \{\alpha^L, \alpha^H\} \to [0,1]^2$ from the current state to a fiscal policy (t_n, g_n). Each citizen's strategy is a mapping $\gamma_C : [0,1]^2 \times [0,1] \to \mathbb{R}_+$ from the observed policy and posterior p_n to a report $\tilde{w}_n$.

Payoffs. Each citizen's per-period payoff is $U^{(\kappa)}(\tilde{w}_n)$ as specified in Stage 2. The elite's per-period payoff is

$$U_E(\alpha_n;\, g_n \mid p_n) \;=\; T(g_n \mid p_n)\,\big[\alpha_n\, g_n + \theta(\sigma)\,(1 - g_n)\big], \tag{18}$$

where $T(g_n \mid p_n) \equiv g_n\, T_1(p_n) + (1 - g_n)\, T_0$ is the belief-dependent tax base, with

$$T_1(p_n) \;\equiv\; \frac{c}{4\,(1 - \kappa\, p_n)}, \qquad T_0 \;\equiv\; \frac{c}{4\,(1 - \kappa\, s)}, \qquad s \;\equiv\; \sigma\,\theta(\sigma). \tag{19}$$

The elite maximizes the discounted payoff stream $\sum_{n=0}^{\infty} \delta^n\, U_E(\alpha_n;\, g_n \mid p_n)$ for discount factor $\delta \in [0,1)$. Lemma 4.1 shows that the stage-game payoff comparison is sufficient for the elite's incentive constraint in every equilibrium configuration.

All parameters $(\kappa, c, \sigma, q^H, q^L)$ are common knowledge. Fig. 3 summarizes the within-period timing.

For any fixed belief p, the per-period payoff U_E is convex in g (it is the product of two affine functions of g). Whether this stage-game property extends to the dynamic optimization depends on whether the continuation value favors interior allocations. Lemma 4.1 below shows it does not. Under provision, the belief-dependent tax base reduces to $T_1(p_n)$, generalizing the static provision tax base (equation (16)) by replacing the known α with the posterior belief p. Under rents it reduces to T_0, coinciding with Eq. (17).

Solution concept. The equilibrium concept is Markov Perfect Equilibrium (MPE) (Maskin and Tirole, 2001), refined by the D1 criterion[12] (Banks and Sobel, 1987). Equilibrium requires *belief consistency* (citizens' posteriors follow the elite's equilibrium strategy via Bayes' rule on path and the D1 criterion off path) and *incentive compatibility* (given these beliefs, each elite type prefers its equilibrium action to imitating the other). Together, these conditions ensure that observed reforms are credible and self-enforcing.

[9] In Besley (2020), the value of public spending is publicly observed each period, so no signaling game arises. Here, the elite's private observation makes its allocation choice informative about fiscal fundamentals.

[10] The specification covers two limiting cases: $q^H = q^L$ yields i.i.d. draws on $\{\alpha^L, \alpha^H\}$ (with $\rho = q^L$), the two-state analog of the i.i.d. benchmark in Besley (2020), and $(q^H, q^L) = (1, 0)$ yields an absorbing fixed-α environment, in which case ρ is exogenously pinned down by the initial draw rather than by the stationary formula.

[11] This is a modeling choice. It is consistent with an overlapping-generations interpretation in which each cohort of citizens inherits κ from its predecessors but has no stake in prior periods. Universalizing over strategies would admit history-contingent plans, including grim-trigger punishments of low-type elites, as equilibrium objects. Beliefs about the fundamental α remain unrestricted: citizens may be arbitrarily forward-looking about future states.

[12] Informally, D1 disciplines beliefs about unexpected actions by asking which type of elite would have the strongest incentive to deviate. If the high-value elite benefits from an unexpected deviation for a strictly larger set of citizen responses than the low-value elite does, D1 requires citizens to attribute the deviation to the high-value elite. In our setting, unexpected provision is therefore interpreted as evidence of high α. Strictly speaking, the dynamic environment studied here is not a canonical sender–receiver signaling game, so our use of D1 is an adaptation. It remains appropriate because the within-period interaction retains signaling structure, with a privately informed elite taking an action that citizens observe before responding, and because the type set is fixed (the state α), so the type-by-type belief comparisons on which D1 rests are well-defined. This is a milder departure than in settings where the type set is endogenous and such comparisons break down. Ekmekci and Kos (2023) discuss the difficulty of applying standard belief-based refinements outside the canonical signaling game.

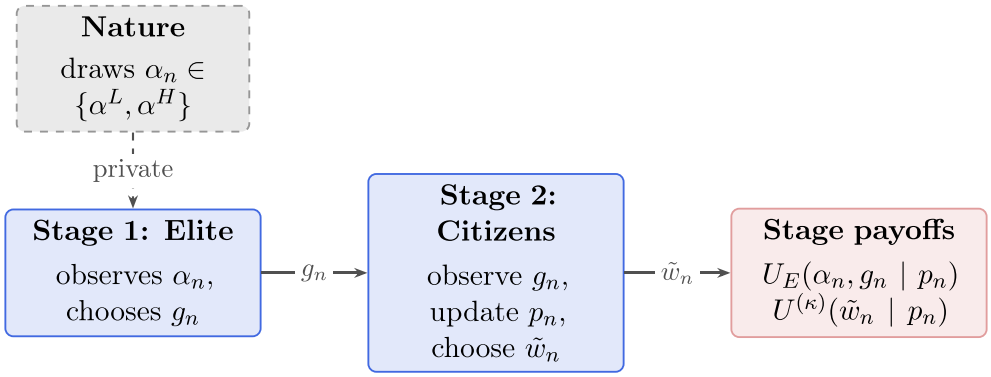


Fig. 3. Timing of the dynamic fiscal signaling game (Definition 4.1). At the start of period n, Nature draws $\alpha_n \in \{\alpha^L, \alpha^H\}$ from a Markov chain with transition probabilities (q^H, q^L). The dashed box and arrow indicate that α_n is privately observed by the elite. The elite then chooses g_n publicly. Citizens observe g_n, update their belief p_n via Bayes' rule (on path) or the D1 refinement (Banks and Sobel, 1987) (off path), and choose $\tilde{w}_n$.

Stage-game payoff gain from provision. The elite's incentive to provide rather than extract is captured by the stage-game payoff difference:

$$\begin{aligned}\Delta(\alpha_n \mid p_n) &= U_E(\alpha_n; g_n{=}1 \mid p_n) - U_E(\alpha_n; g_n{=}0 \mid p_n) \\ &= \frac{c}{4}\left[\frac{\alpha_n}{1-\kappa\, p_n} - \frac{\theta(\sigma)}{1-\kappa\,\sigma\,\theta(\sigma)}\right],\end{aligned} \tag{20}$$

where p_n denotes citizens' belief about α_n. For any given belief p_n, the sign of $\Delta(\alpha_n \mid p_n)$ determines whether provision is optimal for a type with value α_n. The function $\Delta(\alpha_n \mid p_n)$ is strictly increasing in α_n, implying a *single-crossing property*: elite best replies can therefore be represented by a cutoff rule in α_n.

Because citizens' beliefs carry forward across periods, the elite's period-n choice affects future compliance as well as the current payoff. The next lemma shows that this continuation channel does not overturn the within-period incentive, so equilibrium behavior can be characterized by the stage-game payoff gain alone.

Lemma 4.1 (*Stage-game Reduction*)**.** *In any MPE of the game in Definition 4.1, the elite's period-n action depends only on the current state α_n and the contemporaneous citizen belief p_n. The stage-game condition*

$$\Delta\big(\alpha_n \,\big|\, p_n(g_n{=}1)\big) \;\geq\; 0$$

is sufficient for provision to be a best response, and exact whenever future beliefs depend on the period-n action only through the realized state α_n (see the characterization in case 3 of Proposition 4.1).

The proof (Appendix A.3) applies the one-shot deviation principle to each equilibrium configuration, showing that a deviation in g_n either leaves the continuation value unchanged or weakly lowers it. Combined with the convexity of U_E in g_n, the stage-game reduction implies that the corner result $g_n^* \in \{0, 1\}$ established in Section 3 extends to the dynamic game.

4.2. Equilibrium characterization

By Lemma 4.1, equilibrium conditions reduce to stage-game incentive constraints. We seek a separating equilibrium in which $g^*(\alpha^H) = 1$ and $g^*(\alpha^L) = 0$. Under separation, on-path beliefs are correct: citizens infer α^H upon observing $g = 1$ and α^L upon observing $g = 0$, so compliance expands from the rents tax base T_0 to the provision tax base $T_1(\alpha^H)$ in Eq. (19).

The two incentive-compatibility conditions correspond to opposite deviations and must be evaluated at the relevant beliefs.

High-type IC. If the high-value elite deviates from $g = 1$ to $g = 0$, citizens infer α^L. Because the rents tax base T_0 in Eq. (19) does not depend on beliefs about α, the deviation payoff coincides with the on-path comparison at $p = \alpha^H$:

$$\Delta(\alpha^H \mid \alpha^H) \;=\; \frac{c}{4}\left[\frac{\alpha^H}{1-\kappa\,\alpha^H} - \frac{\theta(\sigma)}{1-\kappa\, s}\right] \;\geq\; 0, \tag{21}$$

where $s \equiv \sigma\,\theta(\sigma)$. This yields the lower threshold

$$\kappa \;\geq\; \kappa^H_{\min} \;\equiv\; \frac{\theta(\sigma) - \alpha^H}{\theta(\sigma)\,\alpha^H\,(1-\sigma)}. \tag{22}$$

Low-type IC. If the low-value elite deviates from $g = 0$ to $g = 1$, citizens observe an action that, in the separating candidate, is played on path by the high type with probability one. Bayes' rule therefore delivers $p(g = 1) = \alpha^H$, and the low type's deviation

payoff is $T_1(\alpha^H)\,\alpha^L$, evaluated at the high-type belief,[13] and the IC for separation is

$$\Delta(\alpha^L \mid \alpha^H) \;=\; \frac{c}{4}\left[\frac{\alpha^L}{1-\kappa\,\alpha^H} - \frac{\theta(\sigma)}{1-\kappa\,s}\right] \;\le\; 0, \tag{23}$$

yielding the upper threshold

$$\kappa \;\le\; \kappa^L_{\max} \;\equiv\; \frac{\theta(\sigma)-\alpha^L}{\theta(\sigma)\,\alpha^H - \alpha^L\,s}. \tag{24}$$

Together, (22) and (24) pin down the separating interval $[\kappa^H_{\min}, \kappa^L_{\max})$ within which credible reform is incentive-compatible. The thresholds depend only on the signaling-game primitives $(\alpha^L, \alpha^H, \sigma, c)$ and not on the transition probabilities (q^H, q^L).

When morality is low ($\kappa < \kappa^H_{\min}$), even high-value elites prefer rents, and the equilibrium pools at $g = 0$. When morality is high ($\kappa \ge \kappa^L_{\max}$), low-value elites also find provision attractive and mimic, so separation collapses and the equilibrium pools at $g = 1$. Proposition 4.1 formalizes the three regions and states the additional feasibility condition for pooling at provision.

A fundamentals floor for credible reform. Not all values of α^H above $\sigma\theta(\sigma)$ permit credible reform. Define

$$\underline{\alpha}(c,\sigma) \;\equiv\; \theta(\sigma)\left[\sigma + \tfrac{c}{2}(1-\sigma)\right].$$

When $\alpha^H \le \underline{\alpha}$, the morality threshold $\kappa^H_{\min}$ exceeds the upper bound imposed by Assumption 1, so no admissible κ satisfies the high-type IC and the unique equilibrium is pooling at rents. (The condition $\kappa^H_{\min} \le (1-c/2)/\alpha^H$ rearranges to $\alpha^H \ge \underline{\alpha}$.) The gap $\underline{\alpha} - \sigma\theta(\sigma) = \frac{c}{2}\,\theta(\sigma)(1-\sigma)$ narrows with stronger institutions (higher σ) and lower compliance costs (lower c), so the constraint binds least when the environment is otherwise most favorable to reform. For the remainder of this section, we assume $\alpha^H > \underline{\alpha}$, which is the economically interesting case: the regime in which reform is not materially self-enforcing but morality can restore it.

4.3. Weak high state: morality-activated reform

Consider first the case in which the value of the public good in the high state is only moderately greater than in the low state, but sufficiently above the transfer threshold that morality can activate reform. Formally, assume

$$\alpha^L < \sigma\,\theta(\sigma) < \underline{\alpha}(c,\sigma) < \alpha^H \le \theta(\sigma), \tag{25}$$

so that even in the high state, provision is socially desirable for citizens but not strictly profitable for the elite. This configuration therefore represents a situation in which fundamentals are improving yet still too weak to make reform self-enforcing on material grounds. We refer to it as the *weak-high state* because, although the value of the public good is higher than in the low state, the elite would not provide it voluntarily without moral internalization. In this region, morality is essential for credible reform: the elite provides only if the degree of morality is high enough to make the signal self-enforcing.

Proposition 4.1 (*Morality-activated Reform and Equilibrium Characterization*)**.** *Suppose $\alpha^L < \sigma\,\theta(\sigma) < \underline{\alpha} < \alpha^H \le \theta(\sigma)$, and define the morality thresholds*

$$\kappa^H_{\min} = \frac{\theta(\sigma)-\alpha^H}{\theta(\sigma)\,\alpha^H\,(1-\sigma)}, \qquad \kappa^L_{\max} = \frac{\theta(\sigma)-\alpha^L}{\theta(\sigma)\,\alpha^H - \alpha^L\,\sigma\,\theta(\sigma)}. \tag{26}$$

The equilibrium as a function of morality κ is characterized as follows:

1. ***Pooling at rents*** *($g = 0$). When $\kappa < \kappa^H_{\min}$, even high–α elites find provision unattractive ($\Delta(\alpha^H \mid p = \alpha^H) < 0$). Both types therefore choose $g = 0$, and the equilibrium outcome is pooling at rents, with tax base T_0 given by Eq. (17).*
2. ***Separation (credible reform).*** *If $\kappa^H_{\min} \le \kappa < \kappa^L_{\max}$, there exists a separating equilibrium in which $g^*(\alpha^H) = 1$ and $g^*(\alpha^L) = 0$, and beliefs satisfy $p(g = 1) = \alpha^H$ and $p(g = 0) = \alpha^L$ by Bayes' rule. Upon observing $g = 1$, citizens infer that α is high and update compliance accordingly, leading to an immediate expansion of the tax base from T_0 (equation (17)) to $T_1(\alpha^H)$ (equation (19)).*
3. ***Pooling at provision*** *($g = 1$). When $\kappa \ge \kappa^L_{\max}$, the low-value elite prefers to mimic and separation collapses. A pooling equilibrium with both types providing exists if the high–type incentive constraint under pooling is satisfied. Let the on–path belief be $\bar{\alpha} = \rho\,\alpha^H + (1-\rho)\,\alpha^L$. If $\alpha^H\sigma \le \bar{\alpha}$, pooling at provision obtains whenever $\kappa \ge \max\{\kappa^{pool}, \kappa^{H,\min}\}$, where*

$$\kappa^{pool} = \frac{\theta(\sigma)-\alpha^L}{\theta(\sigma)\,[\bar{\alpha} - \alpha^L\sigma]}, \qquad \kappa^{H,\min} = \frac{\theta(\sigma)-\alpha^H}{\theta(\sigma)\,[\bar{\alpha} - \alpha^H\sigma]}. \tag{27}$$

This threshold is tight when $q^H = q^L$. When $q^H > q^L$, it is an upper bound on the morality required for pool-at-provision, because continuation-value costs from the shifting belief path deter deviation. No pooling equilibrium exists when $\alpha^H\sigma > \bar{\alpha}$.

The separating equilibrium has an immediate fiscal consequence: credible provision triggers a within-period expansion of the tax base.

[13] Evaluating the low-type IC at the prior $p = \alpha^L$ rather than at the Bayesian posterior $p = \alpha^H$ would yield a higher upper threshold and thus a wider separating interval. The Bayesian posterior is therefore the conservative benchmark: it imposes the strongest credibility requirement on the low type. The D1 refinement (Appendix A.2) plays a parallel role off path in the pool-at-rents candidate, where it likewise selects $p(g = 1) = \alpha^H$ and thereby rules out the deviation.

Corollary 4.1 (*Same-period Fiscal Expansion*). *Within the separating region* $\kappa^H_{\min} \le \kappa < \kappa^L_{\max}$, *in the D1-refined separating equilibrium, observing provision* ($g = 1$) *leads citizens to infer* $\alpha = \alpha^H$ *and immediately increase compliance. The tax base rises within the same period from* T_0 *in Eq. (17) to* $T_1(\alpha^H)$ *in Eq. (19). Equivalently, the same-period fiscal multiplier is*

$$J(\kappa;\sigma) = \frac{T_1(\alpha^H)}{T_0} = \frac{1-\kappa\,\sigma\,\theta(\sigma)}{1-\kappa\,\alpha^H} > 1. \tag{28}$$

Corollary 4.2 (*Uniqueness in the Activation Region*). *Suppose* $\alpha^L < \sigma\theta(\sigma) < \underline{\alpha} < \alpha^H \le \theta(\sigma)$ *and* $\kappa \in (\kappa^H_{\min}, \kappa^L_{\max})$. *The separating equilibrium of Proposition 4.1 is the unique D1-refined Markov Perfect Equilibrium of the dynamic fiscal signaling game.*

Morality plays a dual role. First, it activates reform. When $\kappa < \kappa^H_{\min}$, even high-value elites find provision too costly, so both types choose rents and fiscal capacity remains low. Once κ exceeds $\kappa^H_{\min}$ but remains below $\kappa^L_{\max}$, the high-value elite's decision to provide can be credible. In the D1-refined separating equilibrium, the low-value elite refrains from mimicking and the improvement in fundamentals is revealed through behavior. Economically, the key mechanism is that moral preferences can make elites' actions informative about fundamentals, so that citizens update beliefs and compliance immediately when reforms are genuinely warranted. This separating logic is particularly relevant for episodes in which credible public-good provision is interpreted as evidence of improved fiscal fundamentals and triggers a rapid compliance response.

Second, morality amplifies the fiscal response. Within the separating region $[\kappa^H_{\min}, \kappa^L_{\max})$, the same-period increase in the tax base, captured by $J(\kappa;\sigma)$, rises monotonically with κ and exceeds one for any $\kappa > 0$. This reflects the endogenous compliance channel unique to Homo Moralis preferences: citizens internalize the universalized return to provision, which rises with α, so credible reforms trigger an immediate and stronger fiscal expansion than under reciprocity-based mechanisms.

From Eqs. (26), stronger institutions reduce the level of morality required for credible reform. For parameter values such that $\alpha^H > \frac{1}{2}$ and $\alpha^H + \alpha^L > 1$, both morality thresholds $\kappa^H_{\min}$ and $\kappa^L_{\max}$ decrease as institutional quality improves. This implies that when institutions are more cohesive, elites can sustain credible provision with weaker moral support, and citizens can interpret policy actions as more reliable signals of high public-good value.

4.4. Strong high state: fundamentals-driven reform

We now turn to the case in which the value of public goods in the high state is sufficiently large for provision to be privately optimal even in the absence of morality. Formally, assume

$$\alpha^L < \sigma\,\theta(\sigma) < \theta(\sigma) < \alpha^H, \tag{29}$$

so that high-value elites find provision attractive even when $\kappa = 0$. Since $\alpha^H > \theta(\sigma) > \underline{\alpha}$, the admissibility condition $\alpha^H > \underline{\alpha}$ is automatically satisfied. In this region, morality is no longer a precondition for credible provision but continues to amplify its fiscal effects by increasing compliance and expanding the tax base.

Proposition 4.2 (*Strong Provision Incentives and Equilibrium Characterization*). *Suppose* $\alpha^L < \sigma\,\theta(\sigma) < \theta(\sigma) < \alpha^H$. *Then the equilibrium provision pattern as a function of* κ *is as follows:*

1. ***Separation (fundamentals-driven reform).*** *For all* κ *such that*

$$0 \le \kappa < \kappa^L_{\max} = \frac{\theta(\sigma) - \alpha^L}{\theta(\sigma)\,\alpha^H - \alpha^L\,\sigma\,\theta(\sigma)}. \tag{30}$$

   the high-value elite strictly prefers provision while the low-value elite does not. Citizens interpret $g = 1$ *as evidence of high value, forming beliefs* $p(g = 1) = \alpha^H$ *and* $p(g = 0) = \alpha^L$. *Morality reinforces the credibility of this signal but is not required for reform to occur.*

2. ***Pooling at provision*** ($g = 1$). *When* $\kappa \ge \kappa^L_{\max}$, *the low-value elite also finds provision profitable and separation collapses. Both types provide, and* $g = 1$ *becomes uninformative. Let* $\bar{\alpha} = \rho\,\alpha^H + (1-\rho)\,\alpha^L$ *denote citizens' on-path belief. The equilibrium is pooling at provision with tax base*

$$T_1(\bar{\alpha}) = \frac{c}{4\,[1-\kappa\,\bar{\alpha}]}. \tag{31}$$

In the strong-high state, strong provision incentives alone ensure that reform takes place even when $\kappa = 0$, so morality is no longer a precondition for credible provision. Higher levels of morality, however, continue to amplify the fiscal impact of reform by raising compliance and expanding the tax base, consistent with the positive slope of the fiscal multiplier $J(\kappa;\sigma)$ in Eq. (28).

The separating region is narrower than in the weak-high state because the high-value elite's incentive constraint is automatically satisfied, leaving only the low-value elite's constraint binding. As a result, moral preferences play a reinforcing rather than activating role: they magnify the fiscal gains of reform and enhance compliance but are not required for credibility. Morality and strong provision incentives are therefore complementary sources of fiscal capacity, with morality strengthening the revenue response that strong incentives make possible, and pooling at provision obtains for a wider range of morality levels.

4.5. Robustness and extensions

We address the role of equilibrium selection in sustaining separation and sketch an extension in which the average degree of morality in the population is itself an evolutionary outcome.

Equilibrium selection and off-path beliefs. The uniqueness of the separating outcome (Corollary 4.2) depends on citizens' off-path beliefs in alternative equilibrium candidates. The binding case is the pool-at-rents configuration, in which a deviation to $g = 1$ is genuinely off path: without a refinement, pessimistic off-path beliefs $p(g = 1) = \alpha^L$ would sustain pool-at-rents even inside the activation region $[\kappa^H_{\min}, \kappa^L_{\max})$.

Throughout we report D1-refined MPEs to discipline these off-path beliefs. By the single-crossing property of $\Delta(\alpha \mid p)$ in α, the high-value elite benefits strictly more from any off-path deviation to $g = 1$ than the low-value elite does, so D1 uniquely assigns $p(g = 1) = \alpha^H$ (Banks and Sobel, 1987; Cho and Sobel, 1990). In the pool-at-rents candidate, this belief makes deviation profitable for the high type whenever $\kappa > \kappa^H_{\min}$, ruling out pool-at-rents within the activation region.

The refinement is applied stage by stage: D1 disciplines beliefs at each period's signaling subgame. In the pool-at-provision configuration, D1 is silent on the off-path action $g = 0$ because the elite's deviation payoff is belief-independent. We adopt the pessimistic convention $p(g{=}0) = \alpha^L$, which yields the most demanding sufficient condition for pool-at-provision and coincides with the selection a dynamic-D1 argument based on continuation-value differences would produce (Appendix A.3, Case 3). Without imposing a refinement, additional Markov equilibria may be supported by alternative off-path beliefs. The same-period fiscal jump applies in any equilibrium in which provision is interpreted as evidence of the high-value state.[14]

Endogenous morality. The characterization above takes κ as a population parameter. Appendix D shows that a positive level of morality can be sustained endogenously once the within-period signaling game is embedded in a long-run cultural process. Suppose the population consists of *materialists* ($\kappa = 0$) and *universalizers* ($\kappa = \kappa_H > 0$), with share μ_n of universalizers at the start of period n and aggregate morality $\kappa_n \equiv \mu_n \kappa_H$. Because the optimal report in Lemma 2.1 is linear in κ, the within-period characterization of this section carries over with κ_n in place of κ. Each period now ends with a cultural-transmission stage in which agents revise their type sporadically, switching toward whichever type has higher realized per-period utility at the current μ_n (Sethi and Somanathan, 2001). The resulting dynamic is a standard discrete-time replicator (Sandholm, 2010).

Appendix D establishes two results. First, the population converges from any interior initial condition to a stable steady state with $\mu^* \in (0, 1)$ and $\kappa^* = \mu^* \kappa_H > 0$, providing a cultural-evolutionary rationale for the assumption $\kappa > 0$ maintained throughout.[15] Global convergence to an interior μ^* is a distinctive feature of universalization: in the reciprocity dynamics of Besley (2020), by contrast, the cultural state tips to a boundary depending on initial conditions. Second, when the elite's allocation is endogenous, the long-run fiscal regime is determined by where the steady-state morality κ^* falls relative to the separating region of Proposition 4.1: pooling at rents, D1-refined separation, or pooling at provision (Corollary D.1). Cultural evolution pins down the moral share, and the level of κ_H selects the fiscal regime.

5. Discussion

This section situates the model relative to reciprocity-based alternatives, discusses the behavioral and empirical foundations for moral universalization, and examines the theory's scope through historical illustrations and identification strategies.

Relation to reciprocity-based models. The most closely related framework is Besley (2020), in which civic-minded citizens condition compliance on the observed gap between public spending and elite rents. The key distinction is informational. In a signaling environment where fundamentals are private information, reciprocity rewards the observed allocation g while universalization rewards the state α that the allocation reveals. The within-period fiscal multiplier under reciprocity depends on preference parameters but is invariant to inferred fundamentals. Under universalization it inherits them, so credible provision triggers an additional compliance response through belief updating (Appendix C, Remark 1). When the two frameworks are matched on a common enviroment, universalization generates a strictly larger multiplier for any reciprocity preference that does not implicitly perform the inference universalization performs (Lemma C.1). The two frameworks also differ in their cultural dynamics: universalization converges globally to an interior steady state (Section 4.5), whereas reciprocity exhibits tipping and path dependence (Besley, 2020, Proposition 2). Table 1 summarizes.

Foundations for moral universalization. A central premise of the analysis is that citizens exhibit a positive degree of moral universalization ($\kappa > 0$). The evolutionary foundation of Alger and Weibull (2013, 2016) establishes that the evolutionarily stable κ equals the index of assortativity in the matching process (Bergstrom, 2003), which may be close to zero in large anonymous populations. These preferences admit an evolutionary microfoundation but can equally be studied as a preference class in their own right. The mechanism proposed here takes κ as a behavioral primitive, and the foundations developed below do not rely on assortativity.

Recent evidence from moral psychology provides direct support. In a series of experiments, Levine et al. (2020) show that universalization reasoning, asking "what if everyone did this?", is a robust feature of human moral judgment, and Kwon et al. (2023) find that subjects spontaneously apply universalization when evaluating rule violations, judging individually beneficial actions as impermissible when universalizing them would produce bad outcomes. These findings connect to two well-established cognitive biases in psychology: the false consensus effect (Ross et al., 1977), whereby individuals project their own behavior onto others, and magical thinking (Shafir and Tversky, 1992), whereby agents act as if their choices causally influence aggregate outcomes.

[14] The weaker Intuitive Criterion of Cho and Kreps (1987) yields the same selection here. D1 is therefore a conservative choice of refinement.

[15] In the quadratic-cost benchmark, $\mu^* = 1/2$. The precise value reflects the curvature of the misreporting-cost function. The interior, globally stable character of the steady state is robust across specifications.

Table 1
Summary of predictions across behavioral foundations.

	Static (complete information)	Dynamic (asymmetric information)
Selfish agents ($\kappa = 0$)	Standard Laffer curve. Elite provides only if $\alpha > \theta(\sigma)$ (strong provision state). No morality.	No signaling incentive. Pooling at rents unless $\alpha^H > \theta(\sigma)$. Reform occurs only when materially self-enforcing. Compliance is independent of beliefs about α.
Reciprocity (Besley, 2020)	Compliance responds to the spending mix g rather than to the value of public goods α. Civic culture disciplines the elite when the civic-minded share is high.	Compliance responds to observed g rather than to the inferred α, so the within-period fiscal multiplier is invariant to fundamentals (Remark 1). Cultural dynamics tip between high- and low-capacity steady states (Besley, 2020, Proposition 2).
Universalization (this paper)	Moral return $\varphi(g, \alpha, \sigma)$ expands the Laffer curve. Elite provides even when $\alpha \leq \theta(\sigma)$ if $\kappa > \bar{\kappa}(\alpha, \sigma)$ (weak provision state). Morality disciplines the elite.	Morality *activates* credible reform in the weak high state, where provision is not self-enforcing: separation requires $\kappa \geq \kappa^H_{\min}$. The within-period fiscal multiplier is strictly increasing in α^H (Remark 1). Cultural dynamics converge globally to an interior steady state (Section 4.5).

Both generate behavior consistent with $\kappa > 0$ in large anonymous populations without requiring assortative matching, as does the diagnostic reasoning studied by Quattrone and Tversky (1984) in the context of voting.[16]

In experimental economics, the behavioral patterns are consistent with universalization, though not exclusively so. Anonymous public goods games document conditional cooperation (Fischbacher et al., 2001; Fischbacher and Gächter, 2010; Herrmann and Thöni, 2009; Chaudhuri, 2011), a pattern compatible with several behavioral foundations, including selfish preferences under beliefs about others' behavior. Universalization provides a preference-based microfoundation for the same behavioral pattern that does not require beliefs about peer behavior: the citizen conditions on a counterfactual in which others act as she does, rather than on observed peer actions. This distinction matters in atomistic settings such as anonymous tax compliance, where peer reports are unobservable. Such beliefs-based rationalizations lose force in atomistic settings with private actions: Dwenger et al. (2016) document substantial compliance in the tax domain under zero enforcement and without feedback on others' compliance, isolating a pro-social motive. Alger (2025) develops a model of norms and norm change driven by social-Kantian preferences, in which the Kantian component generates unconditional contribution, directly aligned with the universalization reasoning studied here. Most directly, van Leeuwen and Alger (2024) structurally estimate individual-level Kantian morality and report, across 112 subjects, a mean of 0.13 and a median of 0.10, with estimates lying mostly in the interior of $[0, 1]$ (their Table 3). Our main-text specification treats κ as homogeneous across citizens. Since the compliance response is linear in κ, aggregate fiscal capacity in any heterogeneous extension depends only on the population mean (footnote following Lemma 2.1), which is the natural object of comparison.

Finally, the cultural-dynamics extension in Section 4.5 shows that once a positive share of universalizers exists, the trait is sustained endogenously through socialization, with the population converging to a stable interior steady state.

Historical illustrations. The model's most distinctive prediction, the jump-start dynamic in which moral norms enable credible fiscal reform under weak institutions, maps onto specific historical configurations. Two cases illustrate the relevant mechanism.

The Nordic countries are often cited as exemplars of high compliance under moderate penalties. Scandinavian societies sustain unusually high levels of generalized trust and civic participation, rooted in pre-existing equality and impartial government that preceded the construction of modern welfare states (Rothstein and Uslaner, 2005; Uslaner, 2008).[17] The model rationalizes this pattern: when κ is high and public spending is perceived as productive (high α), compliance becomes largely self-enforcing, reducing the marginal role of formal enforcement. This corresponds to the strong provision state in Proposition 3.1, where morality and institutional quality are complements.

Botswana provides a more direct illustration of the jump-start dynamic. At independence in 1966, formal state institutions were minimal, as the British protectorate had invested almost nothing in administrative capacity (Acemoglu et al., 2003). Yet the Tswana had a well-developed tradition of participatory governance through the *kgotla*, a village assembly in which the chief was required to seek consensus before issuing binding decisions (Acemoglu et al., 2003; Robinson and Parsons, 2006). When diamond revenues created a sudden and large increase in the potential value of public spending (a positive shock to α), the post-independence elite channeled these revenues into public goods and infrastructure rather than private extraction, a pattern that persisted for decades and stands in sharp contrast to other resource-rich African states. In the language of the model, Botswana at independence approximated the weak-high state: α had risen sharply, σ remained low, but pre-existing civic norms ($\kappa > 0$) enabled the elite to credibly commit

[16] Salonia (2025) provides an axiomatic foundation for universalization preferences grounded in choice theory, showing that Homo Moralis and Roemer's Kantian equilibrium (Roemer, 2010, 2019) are both special cases of a general class of universalization preferences. Interior κ interpolates between selfish (Nash) and fully universalizing (Kantian) behavior.

[17] A representative cultural marker: in a 1994 SVT interview, Swedish minister Mona Sahlin stated that for a Social Democrat it is *häftigt* ("cool") to pay taxes, adding that "tax is the finest expression of what politics is". The remark has since been widely cited as emblematic of Nordic tax attitudes.

to provision, triggering sustained fiscal expansion. This interpretation is consistent with the emphasis on pre-colonial institutional persistence in Acemoglu et al. (2003) and Robinson and Parsons (2006),[18] but adds a behavioral channel through which civic norms translate into fiscal outcomes via the compliance response.[19]

Identification and the correlation between morality and institutions. A natural concern is that moral universalization and institutional quality may be positively correlated across countries: wherever κ is high, σ may also be high, making it difficult to disentangle the moral channel from institutional enforcement. Three features of the model help address this concern. First, the theory's most distinctive prediction is not about the *level* of compliance but about the *dynamic response* to a shift in fundamentals. The jump-start mechanism predicts that, conditional on weak institutions, societies with stronger moral norms should exhibit larger fiscal responses to improvements in the perceived value of public goods. This conditional prediction generates testable variation even when κ and σ are correlated in the cross section: identification comes from the interaction between moral norms and shocks to α, not from either margin alone. Equivalently, the moral channel is identifiable in the intermediate fundamentals regime: in the strong provision state ($\alpha > \theta(\sigma)$) the elite provides regardless of κ, in the transfer state ($\alpha \leq \sigma\theta(\sigma)$) no admissible morality can rescue provision, and only in the weak provision state ($\sigma\theta(\sigma) < \alpha \leq \theta(\sigma)$) does κ flip the allocation. Empirical tests of the jump-start mechanism should target this intermediate regime, where reform is neither materially self-enforcing nor materially infeasible. Second, the comparative statics in Proposition 3.1 show that when $\alpha > 1/2$, morality and institutions are complements ($\partial\bar{\kappa}/\partial\sigma < 0$): stronger institutions *reduce* the morality threshold required for provision. A positive κ–σ correlation is therefore consistent with the model and does not undermine the moral channel. Rather, it reflects the complementarity the model predicts. Third, the model implies a sharp exclusion restriction at the compliance margin. In the selfish limit $\kappa = 0$, the moral return $\kappa\,\varphi(g, \alpha, \sigma)$ vanishes. With atomistic citizens in a continuum population, the material transfer $\sigma\,\theta(\sigma)(1-g)T$ is taken as given and drops out of the FOC, so conditional on the observed policy (t, g), the citizen's optimal report in Lemma 2.1 is independent of σ. Any cross-country compliance response to σ, holding t, g, and α fixed, therefore operates through moral internalization, providing an empirical handle that separates the two channels even when their levels are correlated. What the model adds is that κ has independent explanatory power *conditional* on σ, a prediction amenable to empirical test using cross-country data on tax morale, enforcement capacity, and fiscal responses to reform episodes.

6. Conclusion

Why do some states sustain high fiscal capacity under weak enforcement while others remain trapped in low-compliance equilibria? This paper offers an answer grounded in moral internalization: when citizens evaluate their compliance decisions under universalized payoffs, the resulting fiscal capacity depends not only on enforcement but also on the degree of moral universalization κ and the effective moral return $\varphi(g, \alpha, \sigma)$, which responds to both the value of public goods α and institutional quality σ. The static model shows that morality expands the feasible tax base and can discipline elites into provision even when material incentives favor rent extraction. The dynamic model identifies a jump-start mechanism through which morality enables credible reform: when the value of public goods increases, sufficiently moral citizens generate a compliance response that makes provision a credible signal, triggering an immediate fiscal expansion. The framework is deliberately parsimonious: a single behavioral parameter κ interacts with standard fiscal primitives to generate these results.

Empirical and policy implications. The theory predicts that tax morale and compliance should be more responsive to improvements in the perceived value of public spending than to changes in deterrence alone, consistent with experimental evidence that compliance responds to the perceived allocation of tax revenues (Alm et al., 1993; Montenbruck, 2023). The jump-start dynamic further predicts that fiscal capacity should respond discontinuously to improvements in the value of public spending in societies with sufficiently strong moral internalization, a pattern testable using reform episodes in countries with varying levels of tax morale. For policymakers, the complementarity between moral internalization and institutional quality suggests that investments in fiscal legitimacy, such as transparent reporting of public spending outcomes, can raise compliance more cost-effectively than enforcement alone. These insights open the door for empirical tests linking survey-based measures of moral universalism, perceptions of government effectiveness, and observed tax capacity.

Directions for future research. We now consider how income heterogeneity might be incorporated into the model. For general income w, the misreporting cost is $C(d) = \frac{1}{2}d^2$ in the deviation $d = \tilde{w} - w$, so it penalizes the amount concealed rather than income itself (see footnote). As a result, the optimal deviation from truthful reporting in Lemma 2.1 is independent of w, and wealth heterogeneity does not generate cross-sectional variation in compliance behavior. This is a tractable benchmark but not the only natural specification. If the misreporting cost were instead proportional to income, for example, $C(d, w) = \frac{1}{2}(d/w)^2$, penalizing the concealment *rate* rather than the concealment *level*, then wealthier citizens would face a lower marginal cost of concealing a given amount, but would also internalize a larger fiscal externality under universalization, since their hypothetical aggregate impact on the

[18] Michalopoulos and Papaioannou (2013, 2015) document the persistence of pre-colonial institutional traits across African ethnicities without separating formal institutional channels from cultural-norm channels. Because formal state capacity was minimal at Botswana's independence, the persistent trait operative in 1966 maps to κ rather than σ in the model.

[19] Papaioannou (2020) plots state capacity against national versus ethnic identification across African countries (Fig. 2 therein). Botswana appears as a positive outlier, with high state capacity and high national identification, consistent with the civic norm channel emphasized here. Papaioannou calls for extending Besley's framework to model fragility traps in low-income countries. Our model provides one such extension, and Section 4.5 (Endogenous morality) sketches how pre-existing civic norms persist endogenously through cultural evolution.

tax base scales with income. This tension could generate non-trivial cross-sectional predictions: under universalization, high-income citizens may comply *more* than selfish models predict, precisely because their universalized counterfactual involves a larger revenue loss. Whether this "pivotality" channel reinforces or attenuates the aggregate capacity effects identified here is an open question that merits further analysis, particularly in light of empirical evidence that compliance behavior and the forms of noncompliance vary systematically across income groups (see, e.g., Slemrod, 2007).

Taken together, the analysis suggests that moral universalization deserves consideration alongside enforcement and reciprocity as a foundation for fiscal capacity, and that its distinctive predictions are amenable to empirical test.

Funding

This work was supported by the Spanish Ministry of Science, Innovation and Universities [grant number PID2022-137382NB-I00].

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

I am grateful to Ingela Alger, Pau Juan-Bartroli, and Enrico Mattia Salonia for insightful comments and suggestions. I also thank the editor, an associate editor, and two anonymous referees for comments that substantially improved the paper.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jebo.2026.107696.

Data availability

No data was used for the research described in the article.

References

Acemoglu, D., 2005. Politics and economics in weak and strong states. J. Monet. Econ. 52 (7), 1199–1226. http://dx.doi.org/10.1016/j.jmoneco.2005.05.001, Political economy and macroeconomics. URL https://www.sciencedirect.com/science/article/pii/S0304393205000838.

Acemoglu, D., Johnson, S., Robinson, J.A., 2003. An african success story: Botswana. In: Rodrik, D. (Ed.), In Search of Prosperity: Analytic Narratives on Economic Growth. Princeton University Press, pp. 80–119.

Algan, Y., Cahuc, P., 2013. Trust and growth. Annu. Rev. Econ. 5 (1), 521–549.

Alger, I., 2025. Norms and norm change-driven by social Kantian preferences.

Alger, I., Dierks, K., Laslier, J.-F., 2025. Does universalization ethics justify participation in large elections? TSE Working Paper 21-1193. Toulouse School of Economics. Revised April 2025.

Alger, I., Laslier, J.-F., 2022. Homo moralis goes to the voting booth: Coordination and information aggregation. J. Theor. Polit. 34 (2), 280–312.

Alger, I., Weibull, J.W., 2013. Homo moralis—preference evolution under incomplete information and assortative matching. Econometrica 81 (6), 2269–2302.

Alger, I., Weibull, J.W., 2016. Evolution and Kantian morality. Games Econom. Behav. 98, 56–67.

Alm, J., Jackson, B.R., McKee, M., 1993. Fiscal exchange, collective decision institutions, and tax compliance. J. Econ. Behav. Organ. 22 (3), 285–303.

Banks, J.S., Sobel, J., 1987. Equilibrium selection in signaling games. Econometrica 55 (3), 647–661.

Bergstrom, T.C., 2003. The algebra of assortative encounters and the evolution of cooperation. Int. Game Theory Rev. 5 (3), 211–228.

Besley, T., 2020. State capacity, reciprocity, and the social contract. Econometrica 88 (4), 1307–1335.

Besley, T., Persson, T., 2009. The origins of state capacity: Property rights, taxation, and politics. Am. Econ. Rev. 99 (4), 1218–1244.

Besley, T., Persson, T., 2011. Pillars of Prosperity: The Political Economics of Development Clusters. Princeton University Press.

Binmore, K.G., 1994. Game Theory and the Social Contract: Just Playing, vol. 2, MIT Press.

Binmore, K.G., 1998. The evolution of fairness norms. Ration. Soc. 10 (3), 275–301.

Chaudhuri, A., 2011. Sustaining cooperation in laboratory public goods experiments: A selective survey of the literature. Exp. Econ. 14 (1), 47–83.

Cho, I.-K., Kreps, D.M., 1987. Signaling games and stable equilibria. Q. J. Econ. 102 (2), 179–221. http://dx.doi.org/10.2307/1885060.

Cho, I.-K., Sobel, J., 1990. Strategic stability and uniqueness in signaling games. J. Econom. Theory 50 (2), 381–413. http://dx.doi.org/10.1016/0022-0531(90)90009-9.

Collier, P., 2017. Culture, politics, and economic development. Annu. Rev. Politi. Sci. 20 (1), 111–125.

Dwenger, N., Kleven, H., Rasul, I., Rincke, J., 2016. Extrinsic and intrinsic motivations for tax compliance: Evidence from a field experiment in Germany. Am. Econ. J.: Econ. Policy 8 (3), 203–232.

Eichner, T., Pethig, R., 2021. Climate policy and moral consumers. Scand. J. Econ. 123 (4), 1190–1226.

Eichner, T., Pethig, R., 2022. Kantians defy the economists' mantra of uniform Pigovian emissions taxes. Ecological Economics 200, 107514. http://dx.doi.org/10.1016/j.ecolecon.2022.107514.

Ekmekci, M., Kos, N., 2023. Signaling covertly acquired information. J. Econom. Theory 214, 105746. http://dx.doi.org/10.1016/j.jet.2023.105746.

Fischbacher, U., Gächter, S., 2010. Social preferences, beliefs, and the dynamics of free riding in public goods experiments. Am. Econ. Rev. 100 (1), 541–556.

Fischbacher, U., Gächter, S., Fehr, E., 2001. Are people conditionally cooperative? Evidence from a public goods experiment. Econom. Lett. 71 (3), 397–404.

Herrmann, B., Thöni, C., 2009. Measuring conditional cooperation: A replication study in Russia. Exp. Econ. 12 (1), 87–92.

Juan-Bartroli, P., 2024. On injunctive norms: Theory and experiment.

Juan-Bartroli, P., Karagözoğlu, E., 2026. Moral preferences in bargaining. Econom. Theory 81, 429–452. http://dx.doi.org/10.1007/s00199-023-01544-7.
Juan-Bartroli, P., Rivero-Wildemauwe, J.I., 2025. Social preferences or moral concerns: What drives rejections in the ultimatum game?. arXiv:2510.22086.
Kotlikoff, L.J., Persson, T., Svensson, L.E., 1988. Social contracts as assets: A possible solution to the time-consistency problem. Am. Econ. Rev. 662–677.
Kwon, J., Zhi-Xuan, T., Tenenbaum, J., Levine, S., 2023. When it is not out of line to get out of line: The role of universalization and outcome-based reasoning in rule-breaking judgments. In: Proceedings of the Annual Meeting of the Cognitive Science Society, vol. 45, p. 45.
Laffont, J.-J., 1975. Macroeconomic constraints, economic efficiency and ethics: An introduction to Kantian economics. Economica 42 (168), 430–437.
van Leeuwen, B., Alger, I., 2024. Estimating social preferences and kantian morality in strategic interactions. J. Politi. Econ. Microeconomics 2 (4), 665–706. http://dx.doi.org/10.1086/732125.
Levine, S., Kleiman-Weiner, M., Schulz, L., Tenenbaum, J., Cushman, F., 2020. The logic of universalization guides moral judgment. Proc. Natl. Acad. Sci. 117 (42), 26158–26169. http://dx.doi.org/10.1073/pnas.2014505117, arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.2014505117.
Lindbeck, A., Nyberg, S., Weibull, J.W., 1999. Social norms and economic incentives in the welfare state. Q. J. Econ. 114 (1), 1–35.
López-Cantor, S., 2025. Individual morality and inequality: Public goods and the political support for redistribution.
Maskin, E., Tirole, J., 2001. Markov perfect equilibrium: I. Observable actions. J. Econom. Theory 100 (2), 191–219. http://dx.doi.org/10.1006/jeth.2000.2785.
Michalopoulos, S., Papaioannou, E., 2013. Pre-colonial ethnic institutions and contemporary african development. Econometrica 81 (1), 113–152.
Michalopoulos, S., Papaioannou, E., 2015. On the ethnic origins of african development: Chiefs and precolonial political centralization. Acad. Manag. Perspect. 29 (1), 32–71.
Montenbruck, L., 2023. Fiscal exchange and tax compliance: Strengthening the social contract under low state capacity. Job Market Paper, University of Mannheim. URL https://lauramontenbruck.github.io/research/fiscal-exchange/jmp_montenbruck.pdf.
Muñoz-Sobrado, E., 2022. Taxing moral agents. CESifo Working Paper No. 9867.
Norman, T.W., 2020. The evolution of monetary equilibrium. Games Econom. Behav. 122, 233–239.
Papaioannou, E., 2020. A comment on: “state capacity, reciprocity, and the social contract” by timothy besley. Econometrica 88 (4), 1351–1358.
Quattrone, G.A., Tversky, A., 1984. Causal versus diagnostic contingencies: On self-deception and on the voter’s illusion. J. Pers. Soc. Psychol. 46 (2), 237–248.
Rivero-Wildemauwe, J.I., 2025. Trade among moral agents with information asymmetries. arXiv preprint arXiv:2505.20551.
Robinson, J.A., Parsons, Q.N., 2006. State formation and governance in botswana. J. Afr. Econ. 15 (suppl_1), 100–140.
Roemer, J.E., 2010. Kantian equilibrium. Scand. J. Econ. 112 (1), 1–24.
Roemer, J.E., 2019. How We Cooperate: A Theory of Kantian Optimization. Yale University Press.
Ross, L., Greene, D., House, P., 1977. The “false consensus effect”: An egocentric bias in social perception and attribution processes. J. Exp. Soc. Psychol. 13 (3), 279–301.
Rothstein, B., Uslaner, E.M., 2005. All for all: Equality, corruption, and social trust. World Politics 58 (1), 41–72.
Salonia, E.M., 2025. A foundation for universalisation in games. CEIS Tor Vergata Research Paper Series, Vol. 23, Issue 5, No. 603.
Sandholm, W.H., 2010. Population Games and Evolutionary Dynamics. MIT Press.
Sarkisian, R., 2017. Team incentives under moral and altruistic preferences: Which team to choose? Games 8 (3), 37.
Sarkisian, R., 2021a. Optimal incentives schemes under homo moralis preferences. Games 12 (1), 28.
Sarkisian, R., 2021b. Screening teams of moral and altruistic agents. Games 12 (4), 77.
Sethi, R., Somanathan, E., 2001. Preference evolution and reciprocity. J. Econom. Theory 97, 273–297.
Shafir, E., Tversky, A., 1992. Thinking through uncertainty: Nonconsequential reasoning and choice. Cogn. Psychol. 24 (4), 449–474.
Slemrod, J., 2007. Cheating ourselves: The economics of tax evasion. J. Econ. Perspect. 21 (1), 25–48.
Tabellini, G., 2010. Culture and institutions: Economic development in the regions of europe. J. Eur. Econ. Assoc. 8 (4), 677–716.
Uslaner, E.M., 2008. Corruption, Inequality, and the Rule of Law. Cambridge University Press.
Weingast, B.R., 1997. The political foundations of democracy and the rule of the law. Am. Political Sci. Rev. 91 (2), 245–263.
Weingast, B.R., 2005. The constitutional dilemma of economic liberty. J. Econ. Perspect. 19 (3), 89–108.