Niclas Berggren*

Trust, Overinclusive Rules, and Selective Compliance

https://doi.org/10.1515/rle-2025-0086
Received November 29, 2025; accepted March 23, 2026; published online April 27, 2026

Abstract: This paper develops a theory of how social trust shapes the relationship between overinclusive formal rules, social norms, and legal enforcement. I first present a static model in which individuals face a rule that forbids an action that is sometimes harmful but often harmless. Individuals differ in prosociality, the share of civic types is interpreted as social trust, and a rule-following norm imposes a violation cost that declines with trust. Higher trust weakens this norm and induces civic types to violate the rule in harmless situations while continuing to obey when violations would be harmful, whereas opportunistic types violate in all states. As a result, higher trust can generate more harmless violations, fewer harmful violations, and higher welfare. I then embed this behavioral block in a policy environment where a government chooses enforcement intensity. Enforcement strengthens the norm but is costly. In equilibrium, low-trust societies select a rigid regime with strict enforcement and full compliance. High-trust societies instead select a flexible regime with low average enforcement intensity. In this regime, officials exercise legal discretion and selectively enforce overinclusive rules mainly in harmful cases, while typically tolerating harmless violations, which generates a stable pattern of tolerated noncompliance in harmless situations. Finally, an overlapping-generations extension endogenizes trust through cultural transmission and yields multiple steady states that differ in trust, enforcement, norms, and the prevalence of harmless versus harmful violations.

Keywords: social trust; social norms; overinclusive rules; selective compliance; selective enforcement; cultural transmission

JEL Classification: D02; D64; K42; Z13

***Corresponding author: Niclas Berggren**, Research Institute of Industrial Economics (IFN), Box 55665, 102 15 Stockholm, Sweden; and Faculty of Business Administration, Prague University of Economics and Business, Winston Churchill Square 4, 130 67 Prague 3, Czechia,
E-mail: niclas.berggren@ifn.se. https://orcid.org/0000-0003-0673-5123

Open Access. © 2026 the author(s), published by De Gruyter. This work is licensed under the Creative Commons Attribution 4.0 International License.

1 Introduction

In many domains of everyday life, formally codified rules are only imperfectly aligned with social expectations and actual behavior. Traffic laws that prohibit jaywalking at empty intersections, prohibitions on cycling on deserted pavements, or bans on minor workplace practices that harm no one are often violated with little social condemnation. Recent work in psychology describes such provisions as "phantom rules": codified rules that are seldom followed and only selectively enforced, with jaywalking as the canonical example (Wylie and Gantman 2023). At the same time, there is substantial cross-country variation in social trust and in everyday rule-following. Surveys and experiments document large differences across societies in trust, civic honesty, and attitudes towards behaviors such as tax evasion, fare-dodging, and other minor rule violations. This paper asks how such differences in trust and in the strength of everyday rule-following norms shape the interaction between overinclusive rules, social norms, and legal enforcement, and how this interaction feeds back into the evolution of trust itself.

I develop a simple model of an overinclusive rule that forbids both harmful and harmless behavior and uses it to study how social trust, norms of rule-following, and legal enforcement interact. In a static baseline, individuals differ in prosociality and the share of civic types is interpreted as social trust. A norm of rule-following imposes a cost on violations that depends negatively on trust. Higher trust then weakens the rule-following norm and induces civic types to selectively violate the rule in harmless states while continuing to obey when violations would be harmful, whereas opportunistic types violate in all states.[1] As a result, higher trust can be associated with more frequent harmless violations of the overinclusive rule, fewer harmful violations, and higher welfare. I then introduce a policymaker who chooses an enforcement intensity for the existing rule and show that the policy problem collapses to a stark choice between a rigid regime with strict enforcement and full legal compliance and a flexible regime with minimal enforcement and selective noncompliance. Finally, I embed this environment in a dynamic model of cultural transmission in which trust evolves endogenously through socialization choices, generating distinct legal and cultural regimes as stable outcomes. The high-trust, flexible enforcement regimes that emerge in the model are reminiscent of what Gelfand et al. (2011) describe as relatively "loose" cultures, which tolerate a wider range of behavior and exhibit weaker sanctioning of minor norm violations. The

1 I use the term "opportunistic" as a convenient label for the non-civic type. The intended meaning is simply that these individuals do not internalize external harms when deciding whether to violate the rule. (Thus, the model is not about opportunistic breach of promises in the contractual sense prevalent in the law-and-economics literature).

constructs are, however, conceptually distinct. In the present framework, social trust reflects expectations about others' prosociality that shape the strength of a rule-following norm and the choice of enforcement regime, whereas tightness and looseness capture the overall strength of normative constraints and sanctioning and are largely treated as exogenous. The two approaches are therefore complementary, and one could study how cross-societal variation in cultural tightness interacts with trust to shape equilibrium patterns of legal regulation and everyday rule-following.

The paper relates to a law-and-economics literature, which finds that overinclusion is often an inherent byproduct of promulgating comparatively simple rules rather than more finely tailored standards. In Kaplow (2013), over- and underinclusiveness arise when a relatively simple rule substitutes for a more fact-sensitive (and typically more costly) standard, so that some socially desirable conduct is deterred and some socially undesirable conduct is insufficiently deterred.[2] More recently, Buccirossi and Immordino (2025) formalize public enforcement with "imperfect" rules, defining overinclusion as prohibiting conduct that has positive external effects in some circumstances and underinclusion as failing to prohibit alternative harmful conduct; the environment studied here corresponds to their overinclusion case. They show that the optimal expected sanction is lower than under a perfect rule not only for overinclusive rules but also for underinclusive rules. Finally, Ehrlich (1982) provides an early economic analysis of optimal enforcement that highlights how low probabilities of apprehension and conviction make realized enforcement outcomes "lottery-like" *ex post*, strengthening the case for tempering sanction severity when distributional and error-avoidance concerns matter. The present paper complements this line of work by focusing on how cross-societal differences in trust and informal rule-following norms can endogenously generate selective compliance with an overinclusive rule and thereby shape the enforcement regime.

The analysis is motivated by the empirical literature on social capital, which documents large cross-country and regional differences in social trust and relates these to institutional performance and economic outcomes. Landmark contributions include Putnam et al. (1993) on civic traditions and institutional quality in Italian regions, Knack and Keefer (1997) and Algan and Cahuc (2010) on the cross-country growth payoff of trust, and Guiso et al. (2004) on the role of social capital in financial development. These and related studies indicate that trust and civic norms are systematically correlated with economic performance and the

2 Kaplow (2013) illustrates the logic with a chemical-discharge example in which a simple rule assigns liability based on average harm, thereby overdetering harmless discharges and underdetering harmful ones relative to a case-by-case standard.

functioning of formal institutions. A complementary literature in finance and political economy, for example Carlin et al. (2009), shows that trust and the legal environment interact in shaping investment and market outcomes.

Closer to the present paper, Aghion et al. (2010) develop a seminal model in which low trust increases political demand for regulation, and regulation in turn crowds out the accumulation of trust, leading to multiple steady states with high regulation and low trust or the reverse. In their framework, regulation and distrust are substitutes in supporting economic exchange, and regulation has dynamic costs because it undermines the learning and accumulation of cooperative attitudes. Relatedly, Bénabou and Tirole (2026) develop a general theory of how laws and social norms interact with intrinsic motivation and reputational concerns, and analyze when the expressive role of law calls for softer or tougher incentives; their framework delivers modified Pigou–Ramsey formulas for optimal incentives and characterizes how legal policy shapes and is shaped by prevailing norms.

The paper also relates to the theoretical literature on law and social norms, especially Posner (2000), Shavell (2002), and Acemoglu and Jackson (2017). Posner (2000) develops an economic theory of social norms in which law can support or undermine norms by altering the payoffs to norm compliance and deviation. Shavell (2002) compares law and morality as alternative regulators of conduct and characterizes the domains in which behavior is optimally governed by morality alone, by law and morality jointly, or by law alone. Acemoglu and Jackson (2017) study the interplay between social norms and legal enforcement in a setting where behavior generates externalities, laws cap behavior, and enforcement relies partly on whistle-blowing; they show that laws that strongly conflict with prevailing norms can backfire and that gradual legal change can shift norms. Relative to these contributions, the present paper puts selective noncompliance with a single overinclusive rule at the center of the analysis and uses social trust as the primitive that shapes both the strength of rule-following norms and the choice between rigid and flexible enforcement regimes.

Finally, the framework connects to recent work in psychology and behavioral law that documents the existence of phantom rules and their motivated enforcement (Wylie and Gantman 2023). That literature emphasizes that violations of certain codified rules are frequent, that enforcement is rare and discretionary, and that such rules are particularly susceptible to being used for targeted punishment. Related empirical work in economics studies norm enforcement in the field and documents when and how third parties intervene to sanction norm violations (Balafoutas and Nikiforakis 2012; Balafoutas et al. 2016). My contribution is to provide a tractable economic model in which the prevalence and enforcement of such rules are jointly determined with social trust and with the structure of legal regulation.

The structure of the paper is as follows. Section 2 develops a static baseline model of an overinclusive rule without explicit legal sanctions, in which individuals differ in prosociality, social trust shapes the strength of a rule-following norm, and higher trust leads to selective noncompliance that increases harmless violations, reduces harmful violations, and raises welfare. Section 3 introduces a policymaker who chooses an enforcement intensity for the rule and shows that the optimal policy is a trust-dependent choice between a rigid regime with strict enforcement and full legal compliance and a flexible regime with minimal enforcement and selective noncompliance. Section 4 embeds this environment in an overlapping-generations model of cultural transmission in which trust evolves endogenously, generating multiple steady states that differ in trust, norms, enforcement regimes, and welfare. Section 5 concludes.

2 A Static Model of Trust, Norms, and Rule Violations

This section introduces the static baseline environment. I study a single overinclusive rule that forbids an action that is sometimes harmful but often harmless, such as crossing at a red pedestrian light. Individuals differ in prosociality: civic types internalize the harm their actions impose on others, opportunistic types do not, and social trust is captured by the share of civic types in the population. The rule has no explicit legal sanction; behavior is instead shaped by a social norm of rule-following that imposes a utility cost on violations, and the strength of this norm is allowed to depend on trust. I characterize individual behavior by type and state, derive the incidence of harmless and harmful violations as a function of trust, and show how welfare responds to changes in trust when the norm is endogenous in this way.

2.1 Environment

Consider a unit mass of individuals. Each individual is of one of two types, *civic* or *opportunistic*. Let $\theta \in (0,1)$ denote the measure of civic types in the population, and $1-\theta$ the measure of opportunistic types. The parameter θ will be interpreted as social trust: higher θ means that a randomly drawn individual is more likely to be civic.

Each individual faces a one-shot decision whether to take an action $a \in \{0,1\}$, where $a = 1$ represents a violation of a formal rule and $a = 0$ represents obedience.

The action $a = 1$ yields a private benefit $b > 0$ to the actor.[3] Examples include crossing a street on a red pedestrian signal or ignoring a minor formal requirement that is costly to comply with.

The consequences of violating the rule depend on a state $s \in \{S, D\}$, where S stands for *safe* and D stands for *dangerous*. Each individual faces a state $s \in \{S, D\}$. The state is drawn independently across individuals from a Bernoulli distribution with

$$\Pr(s = S) = q \in (0, 1), \qquad \Pr(s = D) = 1 - q.$$

The state s is privately observed by the individual before choosing a.

If an individual chooses $a = 1$ in the dangerous state D, the action imposes a negative externality of size $H > 0$ on others. If $a = 1$ in the safe state S, there is no externality. Obeying the rule ($a = 0$) yields zero payoff and no externality in either state.

A formal rule prohibits $a = 1$ in both states. The rule is overinclusive in the sense that it also forbids harmless violations in state S. In this baseline static model, we take the rule as given and abstract from explicit legal sanctions or active enforcement; the rule affects behavior only through a social norm.

2.2 Preferences, Prosociality, and Trust

Types differ in how they internalize the externality from harmful violations.

Assumption 1 (Types and prosociality). Civic types fully internalize the harm they impose on others, while opportunistic types do not internalize any harm. Formally, the material payoff of an individual of type $j \in \{C, O\}$ who takes action a in state s is

$$u_j^{\text{mat}}(a, s) = \begin{cases} 0 & \text{if } a = 0, \\ b - \gamma_j H \cdot \mathbf{1}\{s = D\} & \text{if } a = 1, \end{cases} \tag{1}$$

where $\gamma_C = 1$ and $\gamma_O = 0$.

3 A useful interpretation is that b captures either the net value of engaging in the underlying activity (an activity-intensity margin, such as time saved) or the avoided private cost of compliance (time, hassle, or foregone surplus). The baseline assumes a common b for simplicity. If instead b is heterogeneous and potentially correlated with prosociality, the main effect is to shift who is marginal between compliance and violation and thus the prevalence of benign violations for a given trust environment. The key mechanism continues to apply as long as some relatively civic individuals face intermediate incentives (so they violate only in harmless circumstances) while less civic individuals face stronger incentives to violate more broadly; stronger sorting of incentives by type can attenuate this selective-compliance pattern.

Thus, if a civic agent violates the rule in the dangerous state, she internalizes the resulting harm H as a reduction in her own payoff. An opportunistic agent ignores the harm.

Social trust in this static setting is captured by the population parameter θ. Following Kirby (forthcoming), I use "social trust" in the sense of generalized trust in strangers, namely as proceeding in social interactions on the (default) supposition that others will fulfill their basic obligations, rather than as a claim about their good intentions. In the present environment, the relevant obligations concern refraining from harmful conduct when violations are dangerous, and θ summarizes how prevalent such civic behavior is in the population. Formally, since θ is common knowledge, it is also the probability that a randomly drawn individual will internalize the harm from dangerous violations, which is close in spirit to probabilistic interpretations of generalized trust as a summary assessment of how likely the "average" person is to do the right thing (Coleman 1990; Bjørnskov 2007).[4] In the static model, θ is taken as given. In the dynamic extension, it will be endogenous through cultural transmission.

2.3 The Social Norm of Rule-Following

Beyond intrinsic prosociality, behavior is constrained by a social norm that attaches disutility to violating the formal rule. The strength of this norm depends on the trust environment.

Any violation of the formal rule generates a norm cost $k(\theta)$ borne by the violator, where $k(\theta)$ is decreasing in θ. This norm cost captures both internal discomfort from breaking a widely held rule and the anticipated disapproval or informal sanctions imposed by others. The interpretation is that in low-trust societies, rule violations are taken as signals of dangerous or opportunistic behavior and are strongly condemned, while in high-trust societies rule violations that appear harmless are treated more leniently.

Assumption 2 (Norm strength). Violating the rule ($a = 1$) entails a norm-related cost $k(\theta)$, common to all agents, where $k : (0, 1) \to (0, \bar{k}]$ is continuous and strictly decreasing. Payoff from the norm is

4 Because "civicness" is defined in terms of internalizing harm, θ can also be read as the prevalence of prosocial preferences or intrinsic moral motivation. The model is not designed to distinguish these interpretations at the individual level. By contrast, θ does not capture conditional cooperation or reciprocity in the standard sense, since behavior here does not respond to others' actions or to strategic beliefs about others' compliance. Modeling those concepts would require an explicit interaction structure in which individuals condition their behavior on perceived trustworthiness or on observed past compliance.

$$u^{\text{norm}}(a;\theta) = \begin{cases} 0 & \text{if } a = 0, \\ -k(\theta) & \text{if } a = 1. \end{cases} \tag{2}$$

In this static model, $k(\theta)$ is treated as a reduced-form representation of social expectations and sanctions. A plausible interpretation is that observers update beliefs about an individual's type from observed rule violations, and that social sanctions depend on these beliefs. When civic types are more prevalent, a rule violation is then less informative about opportunism and therefore less stigmatized, which is captured parsimoniously by $k'(\theta) < 0$. I do not model this belief-updating and sanctioning process explicitly, since the analysis only requires the reduced-form dependence of $k(\theta)$ on θ. This reduced-form monotonicity assumption captures an informational "good-faith" interpretation of norm enforcement.[5]

The total payoff of an individual of type j who takes action a in state s is the sum of material and norm-related components,

$$U_j(a, s; \theta) = u_j^{\text{mat}}(a, s) + u^{\text{norm}}(a; \theta). \tag{3}$$

2.4 Individual Behavior

Given θ and the norm strength $k(\theta)$, individuals maximize their own payoff. Throughout, when an individual is indifferent between obeying and violating (i.e. when the net payoff gain from violating is zero), I assume she obeys ($a = 0$). Accordingly, I state violation conditions using strict inequality ($b > k$).

Consider first an opportunistic type. In either state, the opportunistic agent's payoff is

$$U_O(a, s; \theta) = \begin{cases} 0 & \text{if } a = 0, \\ b - k(\theta) & \text{if } a = 1. \end{cases}$$

5 An alternative mechanism, closer to a "broken windows" logic, is that widespread compliance strengthens expectations of literal rule-following, so that deviations become more salient and more strongly condemned, while low-trust environments may exhibit norm breakdown so that violations carry little stigma. Under such a channel, the effective norm cost of violating the formal rule could be increasing in civicness (or more generally non-monotone). If so, higher trust would strengthen informal deterrence of violations, which would shrink (and potentially eliminate) the region in which civic types optimally violate only in harmless states and would mute or reverse the paper's distinctive prediction that higher trust can coincide with more frequent benign violations alongside fewer harmful ones. Allowing the norm cost to depend on the endogenous prevalence of violations is therefore an interesting extension for future work.

Since the externality does not enter his utility, the opportunistic agent does not condition his action on the state s. He violates the rule whenever $b - k(\theta) > 0$ and obeys otherwise.

A civic type conditions on the state s. In the safe state S, her payoff is

$$U_C(a, S; \theta) = \begin{cases} 0 & \text{if } a = 0, \\ b - k(\theta) & \text{if } a = 1, \end{cases}$$

while in the dangerous state D it is

$$U_C(a, D; \theta) = \begin{cases} 0 & \text{if } a = 0, \\ b - H - k(\theta) & \text{if } a = 1. \end{cases}$$

The following result characterizes individual behavior.

Proposition 1 (Behavior by type and state). *Suppose Assumptions* 1 *and* 2 *hold and* $H > b$.

1. *In the dangerous state D, all civic types obey the rule and choose* $a = 0$, *irrespective of* θ.
2. *In the safe state S, civic types choose* $a = 1$ *if and only if* $b > k(\theta)$.
3. *In both states, opportunistic types choose* $a = 1$ *if and only if* $b > k(\theta)$.

Proof. See the Appendix, Proof of Proposition 1. □

Proposition 1 shows that civic types never take harmful actions in the dangerous state. They may, however, violate the rule in harmless situations if the benefit b exceeds the norm cost $k(\theta)$. Opportunistic types either violate the rule in both states or obey the rule in both states, depending on whether the norm is strong enough to deter them.

It is useful to summarize behavior in terms of the thresholds implied by the norm. Let $\hat{\theta}$ denote the unique level of trust (if it exists) at which $b = k(\hat{\theta})$. Since $k(\theta)$ is strictly decreasing, there is at most one such point.

Assumption 3 (Interior region). For expositional simplicity, assume there exists $\hat{\theta} \in (0, 1)$ such that $b = k(\hat{\theta})$. For $\theta < \hat{\theta}$, the norm is strong enough that $b < k(\theta)$; for $\theta > \hat{\theta}$, the norm is weak enough that $b > k(\theta)$.

Under this assumption, behavior takes a simple form.

Corollary 1 (Trust and rule-breaking). *Suppose Assumptions* 1 *and* 2 *hold and* $H > b$. *Under Assumption* 3*:*

1. *If* $\theta < \hat{\theta}$, *then* $b < k(\theta)$ *and all types obey the rule in both states. There are no violations.*
2. *If* $\theta > \hat{\theta}$, *then* $b > k(\theta)$ *and opportunistic types violate the rule in both states, while civic types violate the rule in the safe state and obey in the dangerous state.*

This corollary highlights the basic comparative statics. In low-trust environments, the rule-following norm is strong. Both civic and opportunistic types find it too costly to violate the rule, so they obey even in harmless states. In high-trust environments, the rule-following norm weakens. Opportunistic types resume violating the rule in all states, while civic types selectively violate the rule in harmless situations but continue to obey in situations where violation would be harmful.

2.5 Harmful and Harmless Violations

We can now characterize how the incidence of harmful and harmless violations varies with trust.

Define the following outcome measures:

– The fraction of individuals who violate the rule in the safe state S:

$$v_S(\theta) \equiv \Pr(a = 1, s = S).$$

– The fraction of individuals who violate the rule in the dangerous state D:

$$v_D(\theta) \equiv \Pr(a = 1, s = D).$$

Given independence of types and states, these fractions can be computed as follows.

For $\theta < \hat{\theta}$, all types obey, hence

$$v_S(\theta) = 0, \qquad v_D(\theta) = 0.$$

For $\theta > \hat{\theta}$, Corollary 1 implies that all opportunistic types violate in both states, while all civic types violate only in the safe state. Therefore

$$v_S(\theta) = q \cdot \big[(1-\theta) + \theta\big] = q, \tag{4}$$

$$v_D(\theta) = (1-q) \cdot (1-\theta). \tag{5}$$

Equivalently, the conditional probability of violating in the safe state is $\Pr(a = 1 \mid s = S) = v_S(\theta)/q$, which jumps from 0 to 1 at $\hat{\theta}$.

The following proposition collects the key comparative statics.

Proposition 2 (Trust, harmful and harmless violations). *Suppose Assumptions* 1, 2, *and* 3 *hold and* $H > b$.

1. *The fraction of rule violations in safe states,* $v_S(\theta)$*, is equal to zero for* $\theta < \hat{\theta}$ *and equal to* q *for* $\theta > \hat{\theta}$*. Thus, once trust is high enough that the norm weakens, harmless violations become pervasive.*
2. *The fraction of rule violations in dangerous states,* $v_D(\theta)$*, is equal to zero for* $\theta < \hat{\theta}$ *and equal to* $(1-q)(1-\theta)$ *for* $\theta > \hat{\theta}$*. For* $\theta > \hat{\theta}$*,* $v_D(\theta)$ *is strictly decreasing in* θ*.*

Proof. See the Appendix, Proof of Proposition 2. □

Proposition 2 captures two core implications of the model.

First, once trust is sufficiently high that the rule-following norm weakens, harmless violations of overinclusive rules become widespread. In the example of pedestrian crossings, in a high-trust environment essentially everyone crosses on red when the street is empty.

Second, conditional on being in the regime where the norm has relaxed, greater trust is associated with fewer harmful violations. As θ rises, the share of opportunistic types $(1-\theta)$ shrinks. Since only opportunistic types violate the rule in dangerous states, the incidence of harmful behavior falls with trust.

Taken together, these results deliver the following pattern. High-trust societies can exhibit high levels of rule-violating behavior in situations that are obviously harmless, while simultaneously having few instances of behavior that poses serious risks or harms to others.

2.6 Welfare

To evaluate welfare, I define expected aggregate welfare per capita as private benefits from violations minus norm-related disutility and the external harm H from dangerous violations. Norm-related costs $k(\theta)$ can be interpreted either as pure transfers or as real losses. For simplicity, I treat them as real losses.

For $\theta < \hat{\theta}$, behavior is $a = 0$ for all individuals and states, hence $W(\theta) = 0$. For $\theta > \hat{\theta}$, behavior is as in Corollary 1. In the safe state S, all individuals violate and obtain payoff $b - k(\theta)$. In the dangerous state D, only opportunistic types violate, obtaining $b - k(\theta)$ each and imposing harm H on others. Therefore

$$W(\theta) = q \cdot \big[b - k(\theta)\big] + (1-q) \cdot \big[(1-\theta)\big(b - k(\theta) - H\big) + \theta \cdot 0\big]$$
$$= q\big[b - k(\theta)\big] + (1-q)(1-\theta)\big(b - k(\theta) - H\big). \tag{6}$$

Differentiating (6) with respect to θ for $\theta > \hat{\theta}$ yields

$$\frac{\mathrm{d}W(\theta)}{\mathrm{d}\theta} = -\big[q + (1-q)(1-\theta)\big]k'(\theta) - (1-q)\big(b - k(\theta) - H\big). \tag{7}$$

Since $k'(\theta) < 0$, the term involving $-k'(\theta)$ is positive. The direct term $-(1-q)(b - k(\theta) - H)$ is also positive because $H > b$ and $k(\theta) > 0$. Hence, $\mathrm{d}W(\theta)/\mathrm{d}\theta > 0$ in the high-trust regime.

Proposition 3 (Trust and welfare in the high-trust regime). *Suppose Assumptions* 1, 2, *and* 3 *hold and* $H > b$. *For* $\theta > \hat{\theta}$, *expected welfare* $W(\theta)$ *is strictly increasing in* θ.

Proof. See the Appendix, Proof of Proposition 3. □

The proposition shows that, once the rule-following norm has weakened enough that harmless violations are common, higher trust strictly increases welfare. The mechanism is simple: higher trust reduces the share of opportunistic types, which decreases the incidence of harmful violations in dangerous states, while also slightly weakening the norm and increasing the net surplus from harmless violations.

The static model has three central implications:

(i) Higher trust can lead to less formal rule-following for overinclusive rules, because norms against rule-breaking weaken.
(ii) At the same time, higher trust reduces harmful behavior because the share of opportunistic types falls and civic types never take harmful actions.
(iii) In the high-trust regime, welfare is strictly increasing in trust despite the prevalence of harmless rule violations.

In the next section, a policymaker is introduced who can adjust the formal legal regime, for example by making rules more or less rigid, taking into account these behavioral responses.

3 Regulation and Policy: A Static Extension

In the analysis of Section 2, the formal rule was taken as given and I abstracted from explicit legal sanctions or active enforcement. In the static baseline, there was no explicit legal enforcement. Instead, conditional on intrinsic prosociality, behavior

was governed by a social norm of rule-following, summarized in the norm cost $k(\theta)$. In this section, I introduce a policymaker who can choose an enforcement intensity e for the existing rule. Enforcement both affects the social meaning of violations and hence the norm cost $k(\theta, e)$, and entails a direct resource cost. The behavioral block from Section 2 then describes how individuals respond for any fixed enforcement level. Here, we ask which enforcement regime a welfare-maximizing policymaker would choose given the level of trust θ.

The key trade-off is the following. Strong enforcement increases the normative cost of violating the rule and, in the limit, can induce full legal compliance. This avoids harmful violations in the dangerous state, but also suppresses harmless violations in the safe state and incurs direct enforcement costs. Weak or zero enforcement allows selective noncompliance: civic types violate the rule only in harmless situations, and opportunistic types violate in both states, which creates some harmful behavior but also allows agents to exploit harmless opportunities.

3.1 Enforcement, Norms, and Costs

The policymaker chooses an enforcement intensity

$$e \in [0, \bar{e}],$$

where $\bar{e} > 0$ is an upper bound on feasible enforcement. Think of e as summarizing inspection probability, expected legal penalty, and the general strictness of the enforcement apparatus. I assume enforcement capacity is sufficient to implement the full-compliance benchmark when desired; in particular, the upper bound $\bar{e}$ is large enough to accommodate the compliance threshold derived below.

Enforcement affects the social norm of rule-following. For analytical tractability, assume a simple linear specification.

Assumption 4 (Enforcement and the strength of the rule-following norm). The norm-related cost of violating the rule is

$$k(\theta, e) = k_0 + \alpha e - \beta\theta, \tag{8}$$

where $k_0 > 0$, $\alpha > 0$, and $\beta > 0$. The norm cost enters preferences as

$$u^{\text{norm}}(a; \theta, e) = \begin{cases} 0 & \text{if } a = 0, \\ -k(\theta, e) & \text{if } a = 1. \end{cases} \tag{9}$$

To maintain the interpretation of $k(\theta, e)$ as a (nonnegative) norm-related cost of violating, assume $k_0 > \beta$, so that $k(\theta, e) \geq k_0 - \beta > 0$ for all $\theta \in [0, 1]$ and $e \geq 0$. The parameter β captures the idea from Section 2 that rule violations are more stigmatized in low-trust societies: when θ is low, observed violations are more likely to signal opportunism and are therefore more harshly condemned, so $k(\theta, e)$ is larger and $\partial k(\theta, e)/\partial \theta < 0$. The parameter α captures how strongly enforcement supports the rule-following norm: for given θ, stricter enforcement increases the stigma or psychological cost of violating the rule, so $\partial k(\theta, e)/\partial e > 0$.[6] For given θ and e, the norm-related payoff is identical across types, as in the static baseline.[7]

Enforcement also entails a direct social cost, for example the cost of inspections, policing, and adjudication.

Assumption 5 (Enforcement cost). The enforcement cost is given by a convex function

$$C(e) = \frac{c}{2} e^2, \tag{10}$$

where $c > 0$.

The policymaker does not directly care about fines or transfers associated with enforcement; such transfers are assumed to be redistributed lump-sum and therefore do not affect net welfare. Enforcement affects welfare only through the behavioral responses it induces and through the resource cost $C(e)$.

Individual material payoffs are as in Section 2. With the new norm function, the total payoff of an individual of type $j \in \{C, O\}$ who takes action a in state s at given (θ, e) is

$$U_j(a, s; \theta, e) = u_j^{\text{mat}}(a, s) + u^{\text{norm}}(a; \theta, e), \tag{11}$$

with $u_j^{\text{mat}}(a, s)$ defined in Assumption 1 and $u^{\text{norm}}(a;\ \theta, e)$ as in Equation (9).

---

6 A related literature on motivation crowding studies how formal incentives and controls can weaken individuals' intrinsic willingness to comply with social expectations, e.g., because controls signal distrust or because sanctions reframe a moral obligation as a mere price; see Gneezy and Rustichini (2000); Frey and Jegen (2001); Bénabou and Tirole (2003); Bowles and Polanía-Reyes (2012). In the present specification, enforcement is assumed to strengthen the informal stigma attached to violations, so the within-regime channel is complementary rather than a typical crowding-out mechanism. The parallel is instead at the level of regime choice: when civic motivation and informal constraints are sufficiently strong, additional enforcement has limited marginal deterrence benefits but reduces the scope for benign flexibility under an overinclusive rule, making minimal enforcement optimal (Proposition 5).

7 For $e = 0$, this specification reduces to the baseline norm in Section 2 if we interpret $k(\theta) \equiv k_0 - \beta\theta$, so that $k(\theta, 0) = k(\theta)$. In Sections 3 and 4, I focus on the parameter region $k_0 < b$, which implies that violations are feasible at $e = 0$ for all $\theta \in (0, 1)$ under this linear specialization. Assumption 3 in Section 2 is imposed for expositional clarity in the baseline and is not required for the enforcement analysis.

3.2 Individual Behavior under Enforcement

Given θ and e, behavior is obtained by substituting $k(\theta, e)$ for $k(\theta)$ in the analysis of Section 2. The comparative statics with respect to θ and k are unchanged.

For opportunistic types,

$$U_O(a, s; \theta, e) = \begin{cases} 0 & \text{if } a = 0, \\ b - k(\theta, e) & \text{if } a = 1, \end{cases}$$

independent of the state s. Opportunistic types therefore violate the rule whenever

$$b > k(\theta, e). \tag{12}$$

For civic types, payoffs are

$$U_C(a, S; \theta, e) = \begin{cases} 0 & \text{if } a = 0, \\ b - k(\theta, e) & \text{if } a = 1, \end{cases}$$

in the safe state, and

$$U_C(a, D; \theta, e) = \begin{cases} 0 & \text{if } a = 0, \\ b - H - k(\theta, e) & \text{if } a = 1, \end{cases}$$

in the dangerous state. Under Assumptions 1 and 4 and $H > b$, behavior is therefore exactly as in Proposition 1, with $k(\theta)$ replaced by $k(\theta, e)$. This means that civic types never violate the rule in the dangerous state, opportunistic types violate in both states if and only if $b > k(\theta, e)$, and in the safe state civic types violate if and only if $b > k(\theta, e)$.

This characterization implies a simple threshold in the space (θ, e). Define

$$k(\theta, e) \geq b \quad \Leftrightarrow \quad k_0 + \alpha e - \beta\theta \geq b. \tag{13}$$

If this inequality holds, neither type violates the rule in either state. If it fails, opportunistic types violate in both states and civic types violate in the safe state and obey in the dangerous state.

It is convenient to summarize this threshold in terms of enforcement.

Definition 1 (Enforcement level for full legal compliance). For given θ, define the critical enforcement level

$$e^c(\theta) \equiv \max\left\{0, \frac{b - k_0 + \beta\theta}{\alpha}\right\}. \tag{14}$$

By construction, for any θ with $e^c(\theta) > 0$, we have $k(\theta, e^c(\theta)) = b$.

Corollary 2 (Compliance regimes). *Suppose Assumptions* 1 *and* 4 *hold and* $H > b$.

1. *If* $e \geq e^c(\theta)$*, then* $k(\theta, e) \geq b$ *and all individuals obey the rule in both states. There are no violations.*[8]
2. *If* $e < e^c(\theta)$*, then* $k(\theta, e) < b$*. Opportunistic types violate the rule in both states, and civic types violate the rule only in the safe state.*

Proof. See the Appendix, Proof of Corollary 2. □

Corollary 2 shows that the enforcement choice partitions the space into two regimes. A sufficiently high enforcement level induces full legal compliance. Lower enforcement allows selective noncompliance, where harmless violations in the safe state are tolerated by civic types while opportunistic types violate in both states.

3.3 Welfare as a Function of Enforcement

The policymaker observes θ and chooses enforcement e to maximize expected welfare conditional on θ. Welfare includes private benefits from violations, external harms, norm-related disutility, and the enforcement cost $C(e)$.

As in Section 2, I treat the norm-related cost $k(\theta, e)$ as a real loss. Let $W(\theta, e)$ denote expected welfare per capita net of enforcement costs. Using the same reasoning as in Section 2, for any (θ, e) in the regime with violations, that is when $k(\theta, e) < b$, expected welfare can be written as

$$W(\theta, e) = q\big[b - k(\theta, e)\big] + (1 - q)(1 - \theta)\big[b - k(\theta, e) - H\big] - C(e). \tag{15}$$

The first term is the contribution from the safe state, where all individuals violate and obtain private benefit b but incur norm cost $k(\theta, e)$. The second term is the

8 This "full compliance" regime is a benchmark. It assumes that the enforcement authority can implement sufficiently strict enforcement given the prevailing trust environment and that the determinants of compliance are effectively known. In practice, enforcement is noisy and capacity is bounded by legal, informational, and budget constraints, and policymakers may not observe the underlying trust composition. With imperfect information or limited capacity, one would expect partial deterrence and a smoother transition between regimes, but the central trade-off captured here, avoiding harmful violations versus overdeterring harmless behavior, would remain.

contribution from the dangerous state, where only opportunistic types violate, obtain private benefit b, impose externality H, and incur norm cost $k(\theta, e)$. The last term is the enforcement cost.

If enforcement is strong enough that $k(\theta, e) \geq b$, there are no violations. In that case, welfare simplifies to

$$W(\theta, e) = -C(e), \tag{16}$$

since there is neither private benefit nor harm nor norm cost.

Substituting the linear form of $k(\theta, e)$ from Assumption 4 into (15), write

$$k(\theta, e) = k_0 + \alpha e - \beta\theta, \tag{17}$$

and define

$$A(\theta) \equiv q + (1 - q)(1 - \theta), \tag{18}$$

$$B(\theta) \equiv qb + (1 - q)(1 - \theta)(b - H). \tag{19}$$

Then (15) can be rewritten as

$$\begin{aligned} W(\theta, e) &= B(\theta) - A(\theta)k(\theta, e) - C(e) \\ &= B(\theta) - A(\theta)\big[k_0 + \alpha e - \beta\theta\big] - \frac{c}{2}e^2. \end{aligned} \tag{20}$$

3.4 Optimal Enforcement

The policymaker chooses e to maximize $W(\theta, e)$, taking θ as given and anticipating that behavior is as in Corollary 2. Because behavior is discontinuous at the threshold $e^c(\theta)$, the objective is piecewise defined. On the interval $[0, e^c(\theta))$ we use (20), and on $[e^c(\theta), \bar{e}]$ we use (16).

The next result shows that within the region with violations, welfare is strictly decreasing in enforcement. Hence interior enforcement levels in $(0, e^c(\theta))$ are never optimal.

Lemma 1 (Welfare is decreasing in enforcement when violations occur). *Fix θ and suppose $k(\theta, e) < b$ so that violations occur. Under Assumptions 4 and 5, welfare satisfies*

$$\frac{\partial W(\theta, e)}{\partial e} < 0$$

for all such e.

Proof. See the Appendix, Proof of Lemma 1. □

Lemma 1 implies that, conditional on being in the regime with violations, the policymaker always prefers weaker enforcement. Intuitively, as long as enforcement is not strong enough to eliminate violations altogether, increasing e only raises the norm cost of those violations and adds enforcement costs, without changing behavior.

In the regime with full compliance, welfare is given by (16), and is strictly decreasing in e, since $C'(e) > 0$. Therefore, conditional on targeting full compliance, the policymaker chooses the smallest enforcement level that achieves it.

Combining these observations, the candidate optimal enforcement levels are:

(i) $e = 0$, corresponding to minimal enforcement and selective noncompliance when $k(\theta, 0) < b$;

(ii) $e = e^c(\theta)$, the minimal enforcement level that induces full compliance when $e^c(\theta) > 0$.

Formally:

Proposition 4 (Reduction to two candidate enforcement levels). *For any given θ, the policymaker's optimal enforcement level $e^*(\theta)$ belongs to the set*

$$\{0, e^c(\theta)\}.$$

Proof. See the Appendix, Proof of Proposition 4. □

Proposition 4 reduces the policymaker's problem to a comparison of two regimes: a flexible regime with $e = 0$ and selective noncompliance, and a rigid regime with $e = e^c(\theta)$ and full legal compliance.

Define

$$W^{\text{flex}}(\theta) \equiv W(\theta, 0), \tag{21}$$

$$W^{\text{rigid}}(\theta) \equiv W(\theta, e^c(\theta)). \tag{22}$$

From (20), the flexible regime yields

$$W^{\text{flex}}(\theta) = B(\theta) - A(\theta)\left[k_0 - \beta\theta\right], \tag{23}$$

as long as $k(\theta, 0) = k_0 - \beta\theta < b$, so that violations occur. If instead $k_0 - \beta\theta \geq b$, then even with zero enforcement the norm is strong enough to induce full compliance and[9]

9 Under the parameter restriction $k_0 < b$ used below, this case does not arise; I include it for completeness.

$$W^{\text{flex}}(\theta) = 0. \tag{24}$$

Under the rigid regime, enforcement is set at $e^c(\theta)$, which is the smallest level such that $k(\theta, e^c(\theta)) = b$. In this case there are no violations and welfare is given by (16), that is

$$W^{\text{rigid}}(\theta) = -C(e^c(\theta)) = -\frac{c}{2}\big(e^c(\theta)\big)^2. \tag{25}$$

Substituting the expression for $e^c(\theta)$ from Definition 1 when it is positive,

$$e^c(\theta) = \frac{b - k_0 + \beta\theta}{\alpha}, \tag{26}$$

the rigid regime welfare becomes

$$W^{\text{rigid}}(\theta) = -\frac{c}{2}\left(\frac{b - k_0 + \beta\theta}{\alpha}\right)^2. \tag{27}$$

The policymaker chooses the regime with higher welfare:

$$e^*(\theta) = \begin{cases} 0 & \text{if } W^{\text{rigid}}(\theta) \le W^{\text{flex}}(\theta), \\ e^c(\theta) & \text{if } W^{\text{rigid}}(\theta) > W^{\text{flex}}(\theta). \end{cases} \tag{28}$$

3.5 Comparative Statics of Optimal Enforcement in Trust

To characterize how the optimal enforcement level varies with trust, consider the difference

$$\Delta W(\theta) \equiv W^{\text{flex}}(\theta) - W^{\text{rigid}}(\theta). \tag{29}$$

For parameter values such that the rigid regime is feasible and the norm is not already strong enough to induce compliance at $e = 0$, that is, for θ such that $k_0 - \beta\theta < b$ and $e^c(\theta) > 0$, the explicit expressions (23) and (27) give

$$\Delta W(\theta) = B(\theta) - A(\theta)\big[k_0 - \beta\theta\big] + \frac{c}{2}\left(\frac{b - k_0 + \beta\theta}{\alpha}\right)^2. \tag{30}$$

Under the parameter restrictions in Proposition 5, both regimes are feasible and distinct on a non-empty interval of trust levels. Intuitively, flexible enforcement becomes relatively more attractive as trust rises. A higher share of civic types reduces the expected external damage from opportunistic violations under the flexible regime, while sustaining full compliance in the rigid regime requires increasingly strong (and costly) enforcement because the norm weakens with trust.

Formally characterizing the shape of $\Delta W(\theta)$ in closed form would add little intuition and requires additional inequalities among the parameters. For the remainder of this section, I therefore focus on parameter values such that $\Delta W(\theta)$ is strictly increasing in θ on the interval where $k_0 - \beta\theta < b$ and $e^c(\theta) > 0$, and such that $\Delta W(\theta)$ is negative for low θ in this interval and positive for high θ in this interval. This regularity condition holds on an open set of parameter values (for example, when H is sufficiently large relative to b, k_0, and the enforcement cost parameter c) and captures the idea that as trust rises, the flexible regime yields fewer harmful violations while the rigid regime becomes more expensive to sustain.

These observations imply that, for suitable parameters, there exists a unique cutoff level of trust at which the policymaker is indifferent between the flexible and rigid regimes.

Proposition 5 (Optimal enforcement and trust). *Suppose Assumptions* 1, 4, *and* 5 *hold and $H > b$. Assume parameters satisfy $k_0 < b$. Then there exists a unique trust level $\bar{\theta} \in (0,1)$ such that:*

1. *For $\theta < \bar{\theta}$, the policymaker chooses the rigid regime:*

$$e^*(\theta) = e^c(\theta) > 0,$$

   and full legal compliance is induced.
2. *For $\theta \geq \bar{\theta}$, the policymaker chooses the flexible regime:*

$$e^*(\theta) = 0,$$

   and selective noncompliance occurs.

Moreover, on the region where the rigid regime is optimal, $e^*(\theta)$ is strictly increasing in θ.

Proof. See the Appendix, Proof of Proposition 5. □

Proposition 5 delivers two central comparative statics.

First, for low and moderate levels of trust, the policymaker optimally chooses a rigid enforcement regime that induces full legal compliance. In this region, mistrust is high enough that the expected harm from opportunistic violations justifies the cost of strict enforcement and the welfare loss from suppressing harmless violations. This is consistent with the idea that distrust generates demand for strong regulation and rigid rules, as in Aghion et al. (2010).

Second, for sufficiently high levels of trust, the policymaker optimally chooses a flexible regime with minimal enforcement and selective noncompliance. In this region, the share of opportunistic types is small, so harmful violations are rare.

Allowing citizens to exercise discretion and ignore overinclusive rules in harmless situations yields higher welfare than a regime that enforces full compliance. In equilibrium, high-trust societies can therefore exhibit low levels of formal rule-following for overinclusive rules while sustaining high welfare and low levels of genuinely harmful behavior.

Within the rigid region, the optimal enforcement intensity is increasing in trust, since higher trust weakens the norm of rule-following and more enforcement is required to maintain full compliance. However, the set of trust levels for which strict enforcement is optimal shrinks as overall trust rises. Taken together, the analysis implies that high-trust societies are more likely to operate under flexible regimes with relatively weak enforcement and a socially accepted pattern of selective noncompliance with overinclusive rules.

4 Dynamic Cultural Transmission and the Coevolution of Trust and Enforcement

The previous sections treated social trust θ as an exogenous parameter. In this section, we embed the static environment of Sections 2 and 3 in an overlapping-generations framework where trust evolves endogenously through cultural transmission. In each period, parents decide whether to socialize their children into civic or opportunistic types, taking into account type-specific expected payoffs under the prevailing enforcement regime. The government chooses enforcement myopically period by period. The interaction of these decisions generates dynamics for trust and enforcement, and under natural conditions yields multiple steady states, in the spirit of the cultural-transmission approach in Aghion et al. (2010) but applied here to the coevolution of trust, norms of rule-following, and legal enforcement around an overinclusive rule.

4.1 Overlapping Generations and Timing

Time is discrete, $t = 0, 1, 2, \ldots$, and in each period there is a unit mass of adults who are economically active and a unit mass of children. A fraction $\theta_t \in [0, 1]$ of adults at date t are civic types and $1 - \theta_t$ are opportunistic.

The timeline within a period t is:

1. The government observes θ_t and chooses an enforcement level $e_t \in [0, \bar{e}]$.
2. Adults engage in the static interaction described in Sections 2 and 3, given (θ_t, e_t).
3. Adults become parents and choose how to socialize their children. Cultural transmission is imperfect: parents can bias but not fully determine their

children's types. The distribution of child types in generation $t+1$ is summarized by θ_{t+1}, the share of civic types. The evolution of θ_t depends on the payoff difference between civic and opportunistic types under (θ_t, e_t).

Given the behavior characterized in Sections 2 and 3, period-t welfare is given by the static welfare function $W(\theta_t, e_t)$ defined in Section 3. This overlapping-generations structure isolates the cultural channel: current trust affects behavior and policy today, which in turn affects the incentives to transmit civic traits to the next generation.

4.2 Government Policy in the Dynamic Setting

To keep the dynamics transparent, assume the government is myopic and sets enforcement each period to maximize contemporaneous welfare, taking θ_t as given. That is, in each period

$$e_t \in \arg\max_{e \in [0,\bar{e}]} W(\theta_t, e), \tag{31}$$

where $W(\theta, e)$ is the static welfare function derived in Section 3.

Proposition 5 in Section 3 showed that for given θ, the government's optimal enforcement level belongs to the set $\{0, e^c(\theta)\}$ and depends on θ through a cutoff $\bar{\theta}$. For low trust, strict enforcement is optimal; for high trust, a flexible regime with minimal enforcement is chosen. We can write the optimal enforcement rule as a function $e^*(\theta)$,

$$e^*(\theta) = \begin{cases} e^c(\theta) & \text{if } \theta < \bar{\theta}, \\ 0 & \text{if } \theta \geq \bar{\theta}. \end{cases} \tag{32}$$

where $\bar{\theta}$ is characterized in Proposition 5 and $e^c(\theta)$ is given in Definition 1. At $\theta = \bar{\theta}$ the policymaker is indifferent between the two regimes; by convention, ties are broken in favor of minimal enforcement, so $e^*(\bar{\theta}) = 0$.

Given this policy rule, the pair (θ_t, e_t) evolves in a feedback system where

$$e_t = e^*(\theta_t), \qquad \theta_{t+1} = F(\theta_t), \tag{33}$$

and the key object to be specified is the law of motion $F(\theta)$ for trust.

4.3 Type-Specific Payoffs

For any (θ, e), individual behavior is as characterized in Section 3: civic types never violate the rule in the dangerous state, opportunistic types violate in both states if and only if $b > k(\theta, e)$, and in the safe state civic types violate if and only if

$b > k(\theta, e)$. It is convenient to collect type-specific expected payoffs in the two enforcement regimes identified in that section.

4.3.1 Flexible Regime

When enforcement is sufficiently weak, $e < e^c(\theta)$, the norm cost $k(\theta, e)$ satisfies $k(\theta, e) < b$ and violations occur. The analysis in Section 2 implies that opportunistic types violate the rule in both states, while civic types violate the rule only in the safe state. Expected per-period utilities (including norm costs) are then

$$\pi_C(\theta, e) = q\big[b - k(\theta, e)\big], \tag{34}$$

$$\pi_O(\theta, e) = b - k(\theta, e). \tag{35}$$

The payoff difference is

$$\Delta\pi(\theta, e) \equiv \pi_C(\theta, e) - \pi_O(\theta, e) = (q-1)\big[b - k(\theta, e)\big] = -(1-q)\big[b - k(\theta, e)\big]. \tag{36}$$

In the flexible regime, civic types earn lower per-period payoff than opportunistic types as long as $b - k(\theta, e) > 0$.

4.3.2 Rigid Regime

When enforcement is strong, $e \geq e^c(\theta)$, the norm cost satisfies $k(\theta, e) \geq b$ and, by Corollary 2, all individuals obey the rule in both states. No violations occur, so there is no private benefit b and no external harm. Per-period utilities are

$$\pi_C(\theta, e) = \pi_O(\theta, e) = 0. \tag{37}$$

In the rigid regime, type does not affect behavior in the narrow interaction captured by the model.

The payoff difference $\Delta\pi(\theta, e)$ captures only material and norm-related components in the narrow decision problem around the overinclusive rule. In reality, type may also affect broader lifetime outcomes. Civic individuals may benefit more from living in high-trust environments, for example through better access to cooperative networks or reputational advantages. To capture these forces in reduced form, we introduce a type-specific "cultural payoff" that depends on trust.

Assumption 6 (Cultural payoffs). For each type $j \in \{C, O\}$, there is an additional payoff term $B_j(\theta)$ that depends on θ. The total expected payoff to a type-j individual in an environment with trust θ and enforcement e is

$$U_j(\theta, e) = \pi_j(\theta, e) + B_j(\theta). \tag{38}$$

Civic individuals benefit more from living in a higher-trust society in the sense that the cultural payoff difference $B_C(\theta) - B_O(\theta)$ is strictly increasing in θ. Define the overall payoff difference

$$\Delta U(\theta, e) \equiv U_C(\theta, e) - U_O(\theta, e). \tag{39}$$

A simple example consistent with Assumption 6 is $B_C(\theta) = \varphi\theta$ and $B_O(\theta) = 0$ for some $\varphi > 0$. More generally, one can allow both $B_C(\theta)$ and $B_O(\theta)$ to be increasing in θ, provided that $B_C(\theta) - B_O(\theta)$ is strictly increasing. Note that this restriction on the cultural component does not imply that the overall payoff difference $\Delta U(\theta, e)$ is monotone in θ, because the material component $\pi_C(\theta, e) - \pi_O(\theta, e)$ can vary with θ through the norm cost in the flexible regime.

4.4 Intergenerational Transmission and the Law of Motion for Trust

At the end of period t, each adult chooses how to socialize her child. Let $U_C(\theta_t, e_t)$ and $U_O(\theta_t, e_t)$ denote the expected lifetime payoffs for a child who will be civic respectively opportunistic, assuming that her adult environment will be characterized by (θ_t, e_t).[10]

Following the cultural transmission literature, assume that the probability that a child becomes civic is increasing in the payoff difference $\Delta U(\theta_t, e_t)$. A convenient specification is a discrete-time replicator dynamic:

$$\theta_{t+1} = \theta_t + \lambda\, \theta_t(1 - \theta_t)\, \Delta U(\theta_t, e_t), \tag{40}$$

where $\lambda > 0$ is a speed-of-adjustment parameter. The term $\theta_t(1 - \theta_t)$ vanishes at $\theta_t \in \{0, 1\}$ and is largest when both types are present in substantial fractions, so payoff differences have the greatest impact in mixed populations. For sufficiently small λ and bounded $\Delta U(\cdot, \cdot)$, this dynamic keeps θ_t in [0,1].

10 This assumption keeps the dynamics Markovian. Forward-looking parents who anticipate future changes in θ and e would lead to richer, but also more complex, dynamics.

Combining (40) with the policy rule $e_t = e^*(\theta_t)$ yields a one-dimensional dynamical system

$$\theta_{t+1} = F(\theta_t) \equiv \theta_t + \lambda\, \theta_t(1-\theta_t)\, \Delta U(\theta_t, e^*(\theta_t)). \tag{41}$$

Steady states are fixed points of F, that is, values $\theta^* \in [0,1]$ such that $\theta^* = F(\theta^*)$. From (40), the fixed-point condition is equivalent to

$$\theta^* \in \{0,1\} \quad \text{or} \quad \Delta U(\theta^*, e^*(\theta^*)) = 0. \tag{42}$$

Interior steady states $\theta^* \in (0,1)$ thus correspond to trust levels at which parents are indifferent between socializing children into civic or opportunistic types.

4.5 Existence of Multiple Steady States

Because the optimal enforcement rule $e^*(\theta)$ features a discrete regime switch at $\bar{\theta}$ (Proposition 5), the composite payoff difference $\theta \mapsto \Delta U(\theta, e^*(\theta))$ can in principle be discontinuous at that point. To keep the trust dynamics tractable (and to mimic the smoothing that would arise from enforcement noise, heterogeneity in b, or imperfect observability of θ), I impose the following reduced-form regularity condition directly on this composite function.

Assumption 7 (Sign pattern for payoff differences). The function $\theta \mapsto \Delta U(\theta, e^*(\theta))$ is continuous on [0,1] and satisfies

1. $\Delta U(0, e^*(0)) < 0$,
2. there exist $0 < \theta_L < \theta_H < 1$ such that

$$\Delta U(\theta_L, e^*(\theta_L)) = 0, \qquad \Delta U(\theta_H, e^*(\theta_H)) = 0,$$

and

$$\Delta U(\theta, e^*(\theta)) = \begin{cases} < 0 & \text{for } \theta \in (0, \theta_L), \\ > 0 & \text{for } \theta \in (\theta_L, \theta_H), \\ < 0 & \text{for } \theta \in (\theta_H, 1], \end{cases}$$

3. $\Delta U(1, e^*(1)) < 0$.

This assumption says that in very low-trust and very high-trust environments, opportunistic types yield higher expected payoff than civic types, while in an intermediate trust range the reverse holds. The intermediate region reflects a strong complementarity between trust and the payoff to being civic. Once civic behavior is

sufficiently common, the benefits of being civic outweigh the temptation to behave opportunistically.

Assumption 7 can be microfounded by combining the material payoff differences in (36) and (37) with suitable cultural payoff functions $B_C(\theta)$ and $B_O(\theta)$. For example, in low-trust, rigid regimes, types behave identically in the narrow interaction, so $\pi_C = \pi_O = 0$, and a slight baseline bias in favor of opportunism can make $\Delta U < 0$. In intermediate regions where enforcement is weak and trust is moderate, the network and reputation benefits of civic behavior, captured in $B_C(\theta) - B_O(\theta)$, can dominate the material payoff disadvantage from abstaining in dangerous states, making $\Delta U > 0$. In very high-trust environments where the government optimally selects $e^*(\theta) = 0$ and norms are weak, the material advantage of opportunism in exploiting the remaining safe-state opportunities can again dominate.

Given Assumption 7, the next result characterizes the number of steady states.

Proposition 6 (Existence of multiple steady states). *Suppose Assumptions* 1, 4, 5, 6, *and* 7 *hold. Then the dynamical system*

$$\theta_{t+1} = F(\theta_t) = \theta_t + \lambda\ \theta_t(1-\theta_t)\ \Delta U(\theta_t, e^*(\theta_t))$$

admits four steady states:

1. *the corner steady state* $\theta = 0$*;*
2. *the corner steady state* $\theta = 1$*;*
3. *a low-trust interior steady state* $\theta_L^* = \theta_L \in (0,1)$*;*
4. *a high-trust interior steady state* $\theta_H^* = \theta_H \in (0,1)$*.*

Moreover, these are the only interior steady states.

Proof sketch. See the Appendix, Proof sketch of Proposition 6. □

4.6 Local Stability and Interpretation

Local stability of a steady state θ^* is determined by the derivative of the map F at θ^*. For $\theta^* \in (0,1)$ with $\Delta U(\theta^*, e^*(\theta^*)) = 0$, differentiating (40) yields

$$F'(\theta^*) = 1 + \lambda\ \theta^*(1-\theta^*)\ \frac{\mathrm{d}}{\mathrm{d}\theta}\Delta U(\theta, e^*(\theta))\Big|_{\theta=\theta^*}. \tag{43}$$

For sufficiently small λ and given bounded payoff differences, the sign of $F'(\theta^*) - 1$ is the sign of $\frac{\mathrm{d}}{\mathrm{d}\theta}\Delta U(\theta, e^*(\theta))\Big|_{\theta=\theta^*}$. In particular, if $\frac{\mathrm{d}}{\mathrm{d}\theta}\Delta U(\theta, e^*(\theta))\Big|_{\theta=\theta^*} < 0$, then $F'(\theta^*) < 1$, and for λ small enough it holds that $|F'(\theta^*)| < 1$, which implies local stability.

Assumption 7 implies that $\Delta U(\theta, e^*(\theta))$ has two interior zeros at θ_L and θ_H. These zeros correspond to interior steady states of the trust dynamic, and I denote them $\theta_L^* \equiv \theta_L$ and $\theta_H^* \equiv \theta_H$. Moreover, $\Delta U(\theta, e^*(\theta))$ crosses zero upward at θ_L^* and downward at θ_H^*. In particular,

$$\Delta U'(\theta_L^*, e^*(\theta_L^*)) > 0, \qquad \Delta U'(\theta_H^*, e^*(\theta_H^*)) < 0.$$

Hence, for sufficiently small λ, the low-trust interior steady state θ_L^* is locally unstable and the high-trust interior steady state θ_H^* is locally stable.

Figure 1 illustrates, in panel (a), the implied phase diagram in the canonical replicator style. In the diagram, the vertical object is the one-period change in trust, $\Delta\theta = F(\theta) - \theta$, so steady states correspond to the zeros of $\Delta\theta$. For interior θ, the sign of $\Delta\theta$ coincides with the sign of $\Delta U(\theta, e^*(\theta))$, which determines whether trust rises or falls. The stability result above corresponds to the slope at the crossing. An upward crossing at θ_L^* implies that small deviations to the right raise trust further and small deviations to the left reduce it, whereas a downward crossing at θ_H^* implies local convergence from both sides. Panel (b) of Figure 1 plots the associated policy function $e^*(\theta)$: in the rigid region, maintaining full compliance requires an enforcement intensity $e^c(\theta)$ that increases with trust, while once trust exceeds the cutoff $\bar{\theta}$, the policymaker switches discretely to minimal enforcement (ties at $\bar{\theta}$ broken in favor of minimal enforcement).

The resulting picture is the familiar one from models of cultural and institutional complementarities. There is an unstable low-intermediate trust level θ_L^* that acts as a threshold. Initial conditions with $\theta_0 < \theta_L^*$ converge to a low-trust

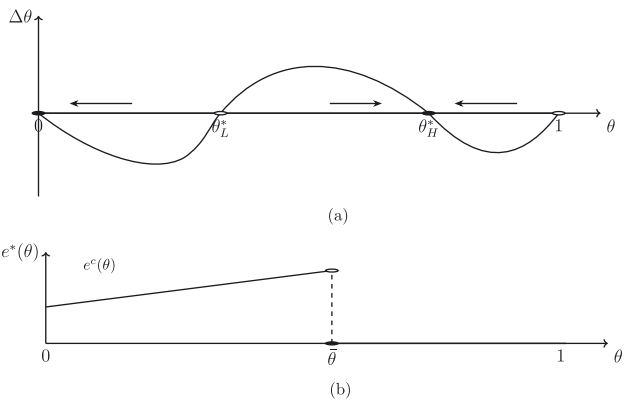

Figure 1: Schematic diagrams. Panel (a) shows the phase diagram for trust under a replicator-type law of motion. Panel (b) shows the associated optimal enforcement policy as a function of trust, with a regime switch at $\bar{\theta}$ (ties broken in favor of minimal enforcement).

region with relatively rigid enforcement and strong rule-following norms. Initial conditions with $\theta_0 > \theta_L^*$ converge to a high-trust steady state θ_H^* associated with low enforcement, weaker rule-following norms, and selective noncompliance with overinclusive rules.

In the low-trust region, opportunistic types are relatively attractive and the government optimally chooses strict enforcement to deter harmful violations. Norms of rule-following are strong and formal compliance is high, but welfare is depressed by the overinclusive rule and the cost of enforcement. In the high-trust region, civic types are attractive, the population becomes predominantly civic, and the government optimally chooses a flexible regime with low enforcement. Norms of rule-following are weaker, harmless violations are common, but harmful behavior is rare and welfare is high.

The dynamic analysis therefore rationalizes the coexistence of different institutional and cultural configurations. Societies with similar fundamentals can settle into either a low-trust, legalistic equilibrium with strong rule-following norms and rigid enforcement, or a high-trust equilibrium with weak rule-following norms for overinclusive rules, selective noncompliance, and high welfare. The static results of Sections 2 and 3 describe behavior and policy in each regime; the OLG extension shows how cultural transmission and policy jointly select between these regimes over time.

5 Concluding Remarks

This paper has asked how social trust shapes the interaction between overinclusive formal rules, social norms, and legal enforcement, and how that interaction feeds back into the evolution of trust. The central message is that higher trust need not manifest itself as stricter formal compliance. When rules are overinclusive, high-trust societies can optimally rely on weaker rule-following norms and more flexible enforcement, so that minor violations of formal rules are common yet harmful behavior is rare and welfare is high. Low-trust societies, by contrast, rationally gravitate toward rigid regulatory regimes with strong rule-following norms, strict enforcement, and high measured compliance, but lower welfare.

Relative to the existing literature, the contribution is threefold. First, the analysis provides a simple mechanism that links social trust to selective noncompliance with overinclusive rules. In the baseline environment, trust determines the strength of a rule-following norm, which in turn governs whether civic individuals obey a rule that sometimes bans harmless behavior. This connects the social capital literature on trust and civicness to the behavioral law and psychology literature on “phantom rules” by showing how high trust can coexist with widespread,

socially accepted violations of certain formal rules. Second, the policy extension characterizes how optimal enforcement depends on trust. In contrast to models where regulation and distrust are substitutes in supporting market exchange (e.g. Aghion et al. 2010), the present framework focuses on the design and enforcement of a single overinclusive rule and shows that trust selects between a rigid, legalistic equilibrium and a flexible, discretionary one. Third, the dynamic extension embeds these static forces in a cultural-transmission framework, generating multiple steady states in which trust, norms of rule-following, and legal regimes are jointly determined. This links the law-and-norms literature (e.g. Posner 2000; Shavell 2002; Acemoglu and Jackson 2017; Bénabou and Tirole 2026) to models of endogenous trust by emphasizing that legal design and trust formation are intertwined rather than separable.

The framework has several implications for how observed compliance and regulatory styles should be interpreted. High rates of formal compliance are not automatically a sign of institutional strength. In low-trust environments, rigid enforcement and strong rule-following norms can be second-best responses to the risk posed by opportunistic behavior, yet they destroy surplus by suppressing harmless actions and absorb resources in monitoring and sanctions. Conversely, lower measured compliance with certain overinclusive rules in high-trust environments need not reflect weak norms or lax institutions. It can be the surface manifestation of a regime that relies on internalized concern for actual harm, informal discretion, and a shared understanding that some rules are effectively focal rather than binding.

The results also highlight that legal and cultural interventions cannot be evaluated in isolation. Policies that mechanically tighten enforcement of overinclusive rules in already high-trust societies may produce little additional deterrence of harmful behavior and can undermine the perceived legitimacy of law if they clash with widely shared informal understandings of acceptable conduct. In low-trust environments, by contrast, efforts to relax enforcement without addressing underlying trust and civicness risk increasing harmful behavior and eroding confidence in the legal system. The dynamic model suggests that durable improvements in trust are more likely to arise from changes in the underlying payoffs to civic behavior, such as higher quality of government, reduced corruption, and denser civic and economic networks, than from simple adjustments to enforcement intensity.

Several limitations are worth noting. First, the norm cost of violating the formal rule is specified in reduced form and is assumed to be lower in higher-trust environments. If instead violations become more salient when compliance is widespread, or if informal sanctions erode when norms break down, some comparative statics could differ. Second, the private gain from noncompliance is assumed common. Allowing it to vary with circumstances or activity intensity, and to correlate with prosociality, would mainly shift who is marginal and thus the prevalence

of benign violations. Third, the policy analysis treats the enforcement authority as able to implement fully deterrent enforcement and to condition policy on the prevailing trust environment. Enforcement noise, bounded capacity, and imperfect information would likely smooth the knife-edge regime switch into partial deterrence, without eliminating the central trade-off between preventing harmful violations and overdeterring harmless behavior. We leave a fuller analysis of these points for future work.

The analysis opens several directions for further work. On the theoretical side, an obvious extension is to abandon the assumption of myopic policy and study optimal enforcement when the government internalizes the effect of current enforcement on future trust, turning the problem into a dynamic policy-commitment or mechanism-design problem. Another avenue would be to introduce multiple domains of behavior with rules that differ in their degree of overinclusiveness and harm prevention, and to investigate how trust shapes the optimal portfolio of regulations and the allocation of enforcement effort across domains. On the empirical side, the model yields concrete predictions: individuals with higher social trust should be more tolerant of harmless violations of overinclusive rules, more demanding of strict sanctions for harmful behavior, and less supportive of rigid, one-size-fits-all regulation; regulatory regimes should covary systematically with both average trust and the distribution of trust across groups. Combining survey measures of trust, attitudes to minor and major legal violations, and detailed regulatory indicators would allow for direct tests of these patterns. Existing experimental work on phantom rules already documents the type of overinclusive, frequently violated provisions and motivated enforcement that the model seeks to rationalize (Wylie and Gantman 2023). Extending such designs across institutional environments with different trust levels and regulatory styles would offer a natural way to test the comparative statics developed here. More broadly, the framework provides a lens for thinking about how societies can move from low-trust, rule-bound equilibria toward high-trust, flexible equilibria, and about the role that legal reforms and institutional improvements can realistically play in supporting that transition.

Acknowledgments: The author wishes to thank Andreas Bergh and Jerg Gutmann for inspiration and two anonymous reviewers for helpfull suggestions. This publication was made possible through the support of Anders Sandrews Stiftelse and of Grant 63522 from the John Templeton Foundation. The opinions expressed in this publication are those of the author and do not necessarily reflect the views of the two financiers.
Conflicts of interest: None to declare.

Appendix: Proofs

Proof of Proposition 1

Proof. Part (1) follows from $H > b$ and $k(\theta) > 0$, which imply $b - H - k(\theta) < 0$. Hence $U_C(1, D;\ \theta) < 0 = U_C(0, D;\ \theta)$, so civics obey in state D. Parts (2) and (3) follow from direct comparison of $U_C(1, S;\ \theta)$ and $U_C(0, S;\ \theta)$ for civics in the safe state, and of $U_O(1, s;\ \theta)$ and $U_O(0, s;\ \theta)$ for opportunists in each state $s \in \{S, D\}$. □

Proof of Proposition 2

Proof. Part (1) follows directly from the expressions above for $v_S(\theta)$. Part (2) follows from (5) and the fact that $1 - \theta$ is strictly decreasing in θ. □

Proof of Proposition 3

Proof. Immediate from (7) and the sign restrictions described above. □

Proof of Corollary 2

Proof. Immediate from (13), Definition 1, and the individual best responses described above. □

Proof of Lemma 1

Proof. For $k(\theta, e) < b$, use (20). Differentiating with respect to e yields

$$\frac{\partial W(\theta, e)}{\partial e} = -A(\theta)\alpha - ce.$$

Since $A(\theta) > 0$, $\alpha > 0$, and $c > 0$, it follows that $\partial W / \partial e < 0$ whenever $e \geq 0$. □

Proof of Proposition 4

Proof. On the interval $[0, e^c(\theta))$, Lemma 1 implies that the maximum of $W(\theta, e)$ is attained at the left endpoint, $e = 0$. On $[e^c(\theta), \bar{e}]$, (16) and Assumption 5 imply that $W(\theta, e)$ is strictly decreasing in e, so the maximum is attained at the left endpoint, $e = e^c(\theta)$. Therefore no value of e outside $\{0, e^c(\theta)\}$ can be optimal. □

Proof of Proposition 5

Proof. The restriction $k_0 < b$ implies $k_0 - \beta\theta \leq k_0 < b$ for all $\theta \in [0, 1]$, and hence $e^c(\theta) > 0$ on (0,1), so the flexible regime with $e = 0$ and the rigid regime with $e = e^c(\theta)$ are feasible and distinct throughout.

By the regularity condition stated in the text, $\Delta W(\theta)$ in (30) is strictly increasing in θ on (0,1). In particular, $W^{\text{flex}}(\theta)$ becomes relatively more attractive as θ rises, because the expected harm from opportunistic violations falls and because maintaining full compliance under the rigid regime requires stricter and more costly enforcement.

For θ close to zero, the share of opportunistic types is close to one, so flexible enforcement produces frequent harmful violations and $W^{\text{flex}}(\theta)$ is low, while rigid enforcement eliminates these harms at a finite enforcement cost, so $\Delta W(\theta) < 0$ for sufficiently low θ. For θ close to one, opportunistic types are rare and the flexible regime yields nearly the full benefit from harmless violations with negligible external harm, while the rigid regime still incurs enforcement costs, so $\Delta W(\theta) > 0$ for sufficiently high θ.

By continuity and strict monotonicity of $\Delta W(\theta)$, there exists a unique $\bar{\theta}$ such that $\Delta W(\bar{\theta}) = 0$. For $\theta < \bar{\theta}$, $\Delta W(\theta) < 0$ and the rigid regime dominates; for $\theta \geq \bar{\theta}$, $\Delta W(\theta) \geq 0$ and the flexible regime dominates (ties are broken in favor of minimal enforcement). The monotonicity of $e^*(\theta)$ on the rigid region follows from the expression for $e^c(\theta)$, which is linear and strictly increasing in θ. □

Proof Sketch of Proposition 6

Sketch of proof. By definition, steady states satisfy $\theta^* = F(\theta^*)$. From (40), this condition is equivalent to

$$\theta^* \in \{0, 1\} \quad \text{or} \quad \Delta U(\theta^*, e^*(\theta^*)) = 0.$$

Assumption 7 states that $\Delta U(\theta, e^*(\theta))$ is continuous, strictly negative on $(0, \theta_L)$, strictly positive on (θ_L, θ_H), and strictly negative on $(\theta_H, 1]$, with zeros at θ_L and θ_H. Hence there are exactly two interior values of θ in (0,1) at which $\Delta U(\theta, e^*(\theta)) = 0$, namely θ_L and θ_H. Together with the corner solutions $\theta = 0$ and $\theta = 1$, this yields the four steady states listed above, and no additional interior steady states exist. A formal f can be obtained by applying the intermediate value theorem and the sign pattern implied by Assumption 7. □

References

Acemoglu, D., and M. O. Jackson. 2017. "Social Norms and the Enforcement of Laws." *Journal of the European Economic Association* 15 (2): 245—95.

Aghion, P., Y. Algan, P. Cahuc, and A. Shleifer. 2010. "Regulation and Distrust." *Quarterly Journal of Economics* 125 (3): 1015—49.

Algan, Y., and P. Cahuc. 2010. "Inherited Trust and Growth." *The American Economic Review* 100 (5): 2060—92.

Balafoutas, L., and N. Nikiforakis. 2012. "Norm Enforcement in the City: A Natural Field Experiment." *European Economic Review* 56 (8): 1773—85.

Balafoutas, L., N. Nikiforakis, and B. Rockenbach. 2016. "Altruistic Punishment Does Not Increase with the Severity of Norm Violations in the Field." *Nature Communications* 7: 13327.

Bénabou, R., and J. Tirole. 2003. "Intrinsic and Extrinsic Motivation." *The Review of Economic Studies* 70 (3): 489—520.

Bénabou, R., and J. Tirole. 2026. "Laws and Norms." *Journal of Political Economy* 134 (2): 731—772.

Bjørnskov, C. 2007. "Determinants of Generalized Trust: A Cross-Country Comparison." *Public Choice* 130 (1—2): 1—21.

Bowles, S., and S. Polanía-Reyes. 2012. "Economic Incentives and Social Preferences: Substitutes or Complements?" *Journal of Economic Literature* 50 (2): 368—425.

Buccirossi, P., and G. Immordino. 2025. "Optimal Sanctions for Imperfect Rules." *American Law and Economics Review*: ahae003. https://doi.org/10.1093/aler/ahae003.

Carlin, B. I., F. Dorobantu, and S. Viswanathan. 2009. "Public Trust, the Law, and Financial Investment." *Journal of Financial Economics* 92 (3): 321—41.

Coleman, J. S. 1990. *Foundations of Social Theory*. Cambridge: Belknap Press of Harvard University Press.

Ehrlich, I. 1982. "The Optimum Enforcement of Laws and the Concept of Justice: A Positive Analysis." *International Review of Law and Economics* 2 (1): 3—27.

Frey, B. S., and R. Jegen. 2001. "Motivation Crowding Theory." *Journal of Economic Surveys* 15 (5): 589—611.

Gelfand, M. J., J. L. Raver, L. H. Nishii, L. M. Leslie, J. Lun, B. C. Lim, et al. 2011. "Differences Between Tight and Loose Cultures: A 33-Nation Study." *Science* 332 (6033): 1100—4.

Gneezy, U., and A. Rustichini. 2000. "A Fine is a Price." *The Journal of Legal Studies* 29 (1): 1—17.

Guiso, L., P. Sapienza, and L. Zingales. 2004. "The Role of Social Capital in Financial Development." *The American Economic Review* 94 (3): 526—56.

Kaplow, L. 2013. “Rules Versus Standards: An Economic Analysis.” In *Scientific Models of Legal Reasoning*, edited by S. Brewer, 11—84. London: Routledge.

Kirby, N. forthcoming. “What is Social Trust?” *Critical Review of International Social and Political Philosophy*, https://doi.org/10.1080/13698230.2025.2528380.

Knack, S., and P. Keefer. 1997. “Does Social Capital Have an Economic Payoff? A Cross-Country Investigation.” *Quarterly Journal of Economics* 112 (4): 1251—88.

Posner, E. A. 2000. *Law and Social Norms*. Cambridge: Harvard University Press.

Putnam, R. D., R. Leonardi, and R. Y. Nanetti. 1993. *Making Democracy Work: Civic Traditions in Modern Italy*. Princeton: Princeton University Press.

Shavell, S. 2002. “Law Versus Morality as Regulators of Conduct.” *American Law and Economics Review* 4 (2): 227—57.

Wylie, J., and A. P. Gantman. 2023. “Doesn’t Everybody Jaywalk? On Codified Rules that are Seldom Followed and Selectively Punished.” *Cognition* 231: 105323.