A Theory of Equality Before the Law*

Daron Acemoglu       Alexander Wolitzky
MIT                      MIT

October 29, 2019

Abstract

We propose a simple model of the emergence of equality before the law. A society can support effort ("cooperation", "pro-social behavior") using the carrot of future cooperation or the stick of coercive punishment. Community enforcement relies only on the carrot and involves low coercion, low inequality, and low effort. A society in which the elite control the means of violence supplements the carrot with the stick, and involves high coercion, high inequality, and high effort. In this regime, elites are privileged by both laws and norms: because they are not subject to the same punishments as non-elites, norms are also more favorable for them. Nevertheless, it may be optimal—even from the viewpoint of the elite—to establish equality before the law, where all agents are subject to the same coercive punishments and norms are more equal. The central mechanism is that equality before the law increases elites’ effort, which in turn encourages even higher effort from non-elites. Equality before the law combines high coercion and low inequality—in our baseline model, elites exert the same level of effort as non-elites. Factors that make the emergence of equality before the law more likely include limits on the extent of coercion, greater marginal returns to effort, increases in the size of the elite group, greater political power for non-elites, and under some additional conditions, lower economic inequality.

Keywords: coercion, community enforcement, inequality, laws, norms, political economy, prisoners' dilemma, repeated games, rule of law.

JEL Classification: P16, P51, K10, C73.

*We thank Bob Gibbons, Gillian Hadfield, Suresh Naidu, Jean Tirole, John Wallis, and seminar participants at Georgetown, MIT, Northwestern, the 2018 NBER Organizational Economics meetings, and the 2019 CIFAR Institutions, Organizations, and Growth meetings for useful discussions and comments. We gratefully acknowledge financial support from the Carnegie Foundation, the NSF, and the Sloan Foundation.
1 Introduction

The notion of equality before the law maintains that laws should apply equally to all citizens: simply put, no one is above the law. This idea—which is also one of the meanings of the amorphous term “rule of law”—is a mainstay of many current constitutions and is widely viewed as a central tenet of a fair and just legal system. Friedrich Hayek saw it as the most critical element of liberal society, stating that “The great aim of the struggle for liberty has been equality before the law” (1960, p. 127). But how and why equality before the law has emerged remains elusive. While some stateless, small-scale societies have egalitarian norms and customs (Bohannan and Bohannan, 1953, Boehm, 1999, Flannery and Marcus, 2014), almost all known historical societies with political hierarchies feature well-defined elites with disproportionate political power—chiefs, kings, lords, military and religious leaders, etc.—as well as laws that privilege these elites. Some scholars, such as Hayek (1960), Jones (1981), and Berman (1983), emphasize the historical roots of equality before the law in Europe, dating back to Greek or Roman legal traditions, the customary laws of Germanic tribes, the English common law tradition, or various turning points in the Middle Ages. More recently, North, Wallis and Weingast (2009) have suggested that equality before the law evolved out of “rule of law among the elites”, meaning a set of practices making all elites subject to the same laws. But why does equality before the law emerge? Specifically, why do elites with disproportionate political and coercive power choose to be bound by the same laws as common citizens?1

This paper provides a simple framework for addressing this question. Our starting point is that society can be organized without a state, relying on norms to support cooperative behavior. Or it can be organized under the auspices of a state with the power to “enforce laws” coercively: that is, to punish deviators. State-enforced laws not only affect incentives directly, but also change norms.2 Coercive state enforcement can exist under “elite domination” (where a subset of agents control the means of violence, enforce laws from which they disproportionately benefit, and are themselves above the law) or under equality before the law (where laws apply equally to all citizens), in each case with different implications for norms, incentives, and inequality. Our main focus is understanding the transition from elite domination to equality before the law.

In our model, a large number of agents repeatedly exert costly effort that generates social benefits. Effort in our model stands for various pro-social behaviors, such as contributions to joint production, public goods, or collective defense, and is perfectly observed. In a society without a

---

1 Of course, equality before the law may be forced on elites. A separate literature (e.g., Rueschemeyer, Stephens and Stephens, 1992, Acemoglu and Robinson, 2000, 2006, Lizzeri and Persico, 2004, Fearon, 2011, Bidner and Françoïs, 2013) studies democratization—how political power shifts from elites towards regular citizens—but does not focus on whether this is accompanied by equality before the law. Our focus is on instances where the elite voluntarily take steps towards equality before the law, though we also discuss how political pressures affect this choice.

2 As emphasized in the context of organizations by Macaulay (1963), Williamson (1985), and Baker, Gibbons and Murphy (1994, 2002), and in the broader context of governance by Granovetter (1985), Ostrom (1990), Milgrom, North, and Weingast (1990), Greif (1993), and Dixit (2003), reputation and the threat of withdrawal of cooperation still matter greatly even when legal enforcement is commonplace.
state, effort can be enforced only by the carrot of continued societal cooperation. This can be achieved by norms supported by standard community enforcement mechanisms. For example, a deviator can be ostracized and excluded from cooperation until she exerts effort to repent for her misdeed. Though community enforcement can support some positive level of equilibrium effort, this level is typically low, owing to the weak nature of this type of enforcement.

An alternative organization of society involves state enforcement of laws, so the carrot of future cooperation is supplemented with the stick of coercive punishment. Here, the means of violence are monopolized by a group of agents (which could be elite agents themselves, law-enforcement officers, or even a band of goons working on behalf of the elite) and can be used to inflict additional punishments on agents who “break the law”. To model an elite-dominated society where some fraction of agents (the elite) are above the law, we assume that elites themselves are not subject to coercion, and we focus on the best equilibrium from their viewpoint.\footnote{To be clear, the elite are above the law but are not “above norms”: when they deviate from equilibrium behavior, they still suffer withdrawal of cooperation. This captures the historical fact that even powerful elites are partially constrained by social norms and the outside options of normal agents. For example, if a feudal lord deviates from social norms, his serfs may flee or rebel. Social norms also regulate cooperation between elites.} In this equilibrium (which we call “elite enforcement”), all agents face the carrot of future cooperation, while normal agents are also confronted with the stick of coercive punishment. Whereas under community enforcement there is relative equality across all agents (in the model, perfect equality), under elite enforcement the threat of punishment makes normal agents work harder than elites, creating inequality. Critically, this inequality is supported by both unequally applied laws and unequal norms—non-elite agents are coercively punished and suffer withdrawal of cooperation from others in society when they do not exert the higher level of effort expected from them. This unequal treatment implies that elites are always better off under elite enforcement, while normal agents may or may not also be better off: they benefit from the higher effort of other normal agents but suffer because they have to exert higher effort than elites.

The core of our analysis asks why and when elites give up the privileges of reduced effort and immunity from coercion that they enjoy under elite enforcement. To do this in the sharpest possible way, we continue to focus on the best equilibrium from the viewpoint of the elite (relaxing this later), but we now also let elites choose to what extent they themselves should be subject to “the law” (i.e., coercive punishments) when they deviate. We establish several key results.

Most importantly, subjecting themselves to coercive punishments has a specific type of commitment benefit for the elite: under equality before the law, because the stick of coercive punishment is used against all agents, the carrot of future cooperation itself becomes more powerful. That is, when elites exert greater effort due to the threat of punishment, the benefits of future cooperation increase, and as a result normal agents are encouraged to work harder as well. This complementarity between coercive enforcement and community enforcement is the key mechanism that may make the elite favor equality before the law.
Table 1 provides a schematic representation of the different enforcement regimes. Community enforcement corresponds to low coercion and low inequality (but also low effort). Elite enforcement involves high coercion and higher effort, but also high inequality favoring the elites. Finally, equality before the law likewise relies on a high level of coercion, but it removes the privileges of the elite and thus involves low inequality (and the highest level of effort from all agents).

<table>
<thead>
<tr>
<th>Low Coercion</th>
<th>High Coercion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low Inequality</td>
<td>Community Enforcement</td>
</tr>
<tr>
<td>High Inequality</td>
<td>?</td>
</tr>
</tbody>
</table>

Table 1: Relationship between enforcement regimes, inequality, and coercion

We also consider implications for social welfare. Greater equality before the law increases both elite and normal agent effort. Under full equality before the law, normal agents are always better off than under elite domination. The utility of the elite themselves may increase (because normal agents exert greater effort) or decrease (because the elite lose their privileged position and are forced to exert greater effort). Finally, under full equality before the law, even from the perspective of the elite themselves, it becomes optimal to have equality of norms as well as laws—the elite give up all of their privileges and exert the same level of effort as normal citizens.

What triggers the transition from elite domination to equality before the law? While our model highlights a number of factors affecting this trade-off, we believe the most important one is the role of violence in society. We show that as the extent of punishments that can be imposed on deviators decreases—for technological, political, or social reasons—it becomes more attractive for the elites to give up their privileges and transition to equality before the law. This follows because, of the two levers affecting normal agents’ incentives, the stick (coercive punishment) becomes less important and the carrot (the promise of future benefits) becomes more important. This changes the trade-off facing the elite and encourages them to increase their own effort. To achieve this increase in own effort, the elite must subject themselves to coercive punishments. This comparative static links our explanation for the emergence of equality before the law to political and social changes, such as the rise of democratic politics (cfr. footnote 1), which strengthen non-elites and put natural limits on how harshly they can be treated by the state or the elite, as well as to social forces limiting

---

4 Table 1 raises the question of whether low coercion and high inequality can be combined. We will return to this question in the context of our model and suggest that the extent of inequality is limited without coercion. Indeed, we are not aware of many historical societies that have combined extreme inequality and low coercion.

We also emphasize that inequality here refers to “inequality of treatment”. Equality of treatment may be associated with inequality of outcomes if skills or other attributes are unequally distributed. See below for comparative statics with respect to inequality of endowments or skills.

5 Of course, when the elite themselves choose to transition to equality before the law, their utility must be greater in this regime than under elite enforcement.

6 In other words, the stick and the carrot are substitutes for the elite. This is because of diminishing marginal returns to effort—the more effort is obtained by coercion, the less valuable is the marginal effort obtained by the threat of withdrawal of cooperation.

---

3
the acceptability of such punishments (e.g., Elias, 1994, Pinker, 2011). Our theory thus gives a novel mechanism by which mass political participation and limits on elites’ power lead to greater equality before the law and associated changes in norms. A complementary comparative static is that if the extent of coercive punishments remains unchanged but the political power of the elite declines (which may again result from the rise of democratic politics), society again moves towards equality before the law.

Equality before the law can also emerge due to factors other than the diminished power of the elite and limits on their ability to impose punishments. A notable possibility is that a change in the nature of production can alter the trade-off facing the elite, for example because effort becomes more important for production or for the provision of vital public goods, such as national defense. However, an overall increase in productivity does not necessarily favor equality before the law because, in addition to increasing the marginal returns to effort, it also increases average returns, and higher average returns encourage elites to maintain their privileges. This comparative static therefore runs counter to simple “modernization” ideas and instead predicts that it is not general increases in prosperity but rather the changing nature of production—specifically, greater marginal product of effort—that contributes to the development of equality before the law.

Many instances of the gradual evolution of equality before the law around the world can be interpreted through the lens of these two comparative static results, and we discuss several historical examples in Section 7. In particular, we illustrate how various limits on coercive punishments triggered a rise of equality before the law using the examples of Athenian civilization circa 6th century BC and Britain in the 19th century. We also discuss how defensive modernization in response to external threats can push society towards equality before the law in the context of the Meiji Restoration in Japan and other 19th century reform movements.

Several other comparative statics are also worth noting. First, greater economic inequality resulting from an increase in elites’ endowments works against equality before the law, because greater endowments discourage elite effort. Second, an expansion of the elite favors equality before the law, because elite privileges start becoming more costly in terms of both foregone production opportunities and negative indirect effects on the effort level of normal citizens. Third, we suppose there are two kinds of elites, of which one can be punished by the other, while the latter group cannot be punished at all. We show that if the political power of the first (weaker) group increases, this favors the emergence of equality before the law. These last two comparative statics provide ways in which the expansion of rule of law among the elite subsequently encourages the broader expansion of equality before the law, as argued by North, Wallis and Weingast (2009). Finally, in another extension, we establish that a shift of political power from low-productivity to high-productivity elites (perhaps approximating the increased political power of commercial interests at the expense of traditional landowners) also favors equality before the law. This comparative static is in line with the historical role of the strengthening of commercial interests in eroding the
privileges of the landowning classes in Europe (e.g., Moore, 1966, Aston and Philpin, 1987).

In addition to the literatures on the historical origins of rule of law and democratization mentioned above, four others need to be highlighted. The first is the literature pioneered by North and Weingast (1989), which interprets constitutions and other institutional features as commitment devices for respecting other groups’ property rights, and thus encouraging greater investment and economic participation.\footnote{Other contributions in the same vein include Levi (1989), Weingast (1997), Acemoglu and Robinson (2000), Myerson (2008), Besley and Persson (2011), and Gehlbach and Keefer (2011).} This insight is closely related to the incomplete contracts approach to organizations (e.g., Williamson, 1975, Grossman and Hart, 1986, Hart and Moore, 1990), where manipulating residual control rights within an organization strengthens some agents’ investment incentives by reducing holdup. The result that equality before the law encourages normal citizens to exert effort by removing elite privileges and increasing elite effort bears some similarity to these insights, but with several important differences. First, equality before the law is not a commitment to a constitutional provision but an alternative organization of society leading to a different repeated game equilibrium. Second, equality before the law affects incentives not by preventing ex post expropriation but by encouraging greater elite effort, which increases the value of future cooperation for normal citizens. Equally important, the two models predict different comparative statics: in the simplest interpretation of North and Weingast, an increase in the elites’ ability to expropriate normal citizens should lead to a greater commitment to property rights (to counteract a stronger temptation to expropriate), while our central result is that an increase in the elites’ ability to punish deviators leads to less equality before the law (as the threat of punishment and the promise of cooperation are substitutes in providing incentives).


Third, we also contribute to the nascent literature on the interplay of laws and norms (Benabou and Tirole, 2011, Aldashev et al., 2012, Acemoglu and Jackson, 2018, Jackson and Xing, 2019). Again, this literature has not previously analyzed elite agents’ incentives to submit to legal equality.

Finally, our paper is related to a number of works emphasizing the dual role of violence in enforcing property rights and predation (Moselle and Polak, 2001, Bates, Greif, and Singh, 2002, Grossman, 2002, Konrad and Skaperdas, 2012). As in this literature, in our model violence incentivizes production, but elites control the means of violence and are privileged. The key mechanism that equality before the law enhances community enforcement does not arise in this literature.
The rest of the paper is organized as follows. Section 2 introduces our baseline environment. Section 3 characterizes the best equilibrium under community enforcement (without the state). Section 4 analyzes the same environment under elite enforcement, while Section 5 studies the optimal degree of equality before the law from the viewpoint of the elite. Section 6 presents our main comparative static results, which delineate factors that encourage the emergence of equality before the law. We discuss several historical examples illustrating our comparative static results in Section 7. Section 8 generalizes the baseline environment to a matching model in which, in addition to benefiting society at large, effort generates private benefits for one’s partner. While in the baseline model elites are privileged only because they exert lower effort than others, in this extended environment the best equilibrium from the viewpoint of elites also involves normal agents working harder when they match with elites. Section 9 extends the model to study within-elite heterogeneity in productivity and the implications of a hierarchical structure within the elite. Section 10 concludes. All proofs are presented in the Appendix.

2 Environment

We consider a simple repeated game model of cooperation in which pro-social behavior can be enforced by both withdrawal of cooperation and coercive punishment.

2.1 The Baseline Environment

There is a continuum of infinitely-lived agents that discount the future with discount factor $\delta \in (0, 1)$. Fraction $\alpha$ of the population are elites, and fraction $1 - \alpha$ are normal. At the beginning of every period, each player $i$ chooses a level of cooperation (effort) $x_i \in \mathbb{R}_+$. When the distribution of effort levels among normal agents is given by $F_N$, the distribution of effort levels among elites is given by $F_E$, and player $i$ exerts effort $x_i$, player $i$’s payoff is

$$(1 - \alpha) \mathbb{E}_{F_N} [f_N (x)] + \alpha \mathbb{E}_{F_E} [f_E (x)] - x_i.$$ 

Here, $f_N$ and $f_E$ are the “benefit production functions” that map units of disutility of effort to units of benefits for society. They are strictly increasing, strictly concave, and bounded, and satisfy $f_N (0) = f_E (0) = 0$ and $f_N’ (0), f_E’ (0) > 1/\delta$. The assumption that $f_N’ (0), f_E’ (0) > 1/\delta$ (and hence $f_N’ (0), f_E’ (0) > 1$) implies that the stage game is a continuous-action version of the prisoners’ dilemma. We allow the functions $f_N$ and $f_E$ to differ for normal and elite agents as these agents may have different roles in production—for example, “effort” by elites could simply correspond to “not expropriating others” (see footnote 14 below), or it could represent business investment.

---

8Effort $x_i$ can be interpreted as general cooperative behavior, contributions to collective action or public goods (including collective defense), or effort directed at production that indirectly benefits other agents.
while normal agents’ effort corresponds to supplying labor. None of our results require these two functions to differ—the key difference between normal and elite agents is their vulnerability to coercion, not their production technologies.\footnote{Throughout, we simplify the analysis by ruling out direct transfers between agents or groups. Footnote 13 explains why allowing such transfers does not change our main results.}

We also assume that effort levels are observed by all agents. This perfect monitoring assumption simplifies the analysis and makes the intuition for our results more transparent.\footnote{Combining a continuum population and perfect monitoring/observability raises measurability issues that make formally defining strategies complicated. Rather than addressing these issues formally, we simply assert that our model is obviously the limit of a large finite population. Indeed, the only reason we assume a continuum rather than a finite population is to ensure that, for both a normal agent and an elite agent, the fraction of other agents with elite status is $\alpha$. Assuming a large finite population and allowing this fraction to differ for normal and elite agents leads to more cumbersome notation without yielding any substantive implications.}

At the end of every period, coercive punishments can be inflicted by a “centralized state” on any subset of agents. The state is not a player in the game and has no preferences—its punishment strategy can be specified freely as part of the description of an equilibrium. The key difference between normal and elite agents is that they differ in their vulnerability to state punishment. If a normal agent is punished by the state, she suffers a disutility of $g \geq 0$. On the other hand, if an elite agent is punished by the state, she suffers a disutility of only $\rho g$, where $\rho \in [0, 1]$ is a parameter measuring the vulnerability of elites to coercive punishment.

In this formulation, the parameter $g$ is a measure of the effective coercive capacity of the state. This coercive capacity depends on state institutions (does the state have the infrastructural power to detect deviators and inflict punishments on them once they are caught?), on unequal access to the means of coercion between the elite and normal agents (are normal agents able to resist punishment?), on the distribution of political power (can normal agents mobilize against harsh punishments?), and on a society’s values (is it socially acceptable to impose harsh punishments on law-breakers?). Meanwhile, the parameter $\rho$ is an inverse measure of the extent to which elites are above the law. When $\rho = 0$, elites are completely above the law and immune to coercive punishment, and as a result they can be incentivized only by the threat of withdrawal of cooperation. When $\rho = 1$, elites are subject to the full force of the law, and like normal agents they can be incentivized by the threat of coercive punishment as well as withdrawal of cooperation. Intermediate values of $\rho$, in turn, represent imperfect levels of equality before the law. Such intermediate values may result in practice either because the elite’s privileges protect them from the full force of legal punishments, or because elites are subject to punishment in some domains but not in others (e.g., they can be punished for murder, but not for mistreating their servants).

Throughout, our equilibrium concept is stationary, symmetric, subgame perfect equilibrium. By symmetry, we mean that all normal agents and all elite agents use the same strategies. By stationarity, we mean that there is a single pair of effort levels $(x, y)$ such that, along the equilibrium path, normal agents exert effort $x$ and elite agents exert effort $y$ in every period.\footnote{Non-stationary equilibria can potentially improve on stationary equilibria in discounted repeated games with}
2.2 A Random Matching Interpretation

The economy described so far is centralized in two ways: each individual’s effort directly benefits everyone in society, and a centralized state directly allocates punishments. We remark that it is straightforward to give a mathematically equivalent decentralized interpretation.

Suppose first that effort is still a pure public good, but the means of coercion are controlled by the elite. Players randomly match in pairs, and an elite agent can punish her partner in the match. Suppose also that punishing one’s partner is costless (so a player is indifferent as to whether or not to punish her partner), and that punishment inflicts disutility $g/\alpha$ on a normal agent and $pg/\alpha$ on an elite (this scaling by $1/\alpha$ keeps the expected disutility of punishment fixed at $g$, as there are $\alpha$ elites in the population). This variant of the model where punishments are carried out by elites is completely equivalent to the baseline model.

Next, suppose that benefits are also generated within matches, and a player only benefits from the effort of her partner. Provided that effort levels are chosen before players observe their partners’ status as normal or elite (while status is subsequently observed at the punishment stage), each agent must choose the same level of effort regardless of her partner’s status, and therefore her effort generates the same expected benefit for everyone. This anonymous matching model thus endogenously generates the pure public good feature that was assumed in the centralized model. This version of the model—where all interactions occur within matches—remains mathematically equivalent to the baseline model. In Section 8, we study a variant of this model where matching is non-anonymous, so players know their partners’ status when choosing effort. In this case, effort is no longer a pure public good, but we will see that our most important results continue to apply.

3 Community Enforcement

We start by characterizing the equilibrium in this environment without coercive punishments. In this case, effort will be encouraged by “norms” of cooperation incentivized by community enforcement. These types of norms are relevant for many societies with limited access to coercive punishments, and they remain important even when such punishments are possible.

3.1 Optimal Community Enforcement Equilibria

Suppose that $\alpha = g = 0$, so that all agents are identical and no coercive punishments are available. This gives a model of community enforcement of cooperation. The following result is standard: for perfect monitoring (e.g., Abreu, 1986). Our objective here, however, is to compare optimal stable social arrangements under different enforcement regimes, which makes non-stationary equilibria difficult to interpret. Another way of motivating stationarity is to note that, due to the concavity of the benefit functions $f_N$ and $f_E$, the ergodic distribution of any non-stationary equilibrium is Pareto-dominated by a stationary equilibrium, so stationarity is without loss from the perspective of undiscounted “long-run welfare”. 
this result, and throughout the paper, we denote the first-best (surplus-maximizing) normal agent effort level by

\[ x^{FB} = \left(f'_N\right)^{-1}(1). \]

**Proposition 1** Under community enforcement, the effort level in every Pareto optimal equilibrium is given by \( x^{CE} = \min\{\bar{x}^{CE}, x^{FB}\}, \) where \( \bar{x}^{CE} \) is the unique positive solution to the equation

\[ x = \delta f_N(x). \] (1)

The intuition is that a player who deviates can save an effort cost of \( x \), but loses a benefit of \( f_N(x) \) in the next period. This loss could be supported by grim trigger strategies, in which cooperation completely breaks down following a deviation. With these strategies, a player’s (per-period) equilibrium payoff is \( f_N(x) - x \), while her best payoff from deviating is \( (1 - \delta) f_N(x) \). Equating the two yields (1).

Grim trigger strategies are one way of supporting the unique optimal equilibrium effort level characterized in Proposition 1, but not the only one. In a different optimal equilibrium, a player’s punishment for deviating in period \( t \) is that in period \( t + 1 \) she must play \( x_i = x \) while her opponents all play \( x_j = 0 \), and all players restart the original equilibrium in period \( t + 2 \) if this punishment is successfully carried out. Relative to grim trigger, this “repentance equilibrium” has the advantage that it is renegotiation proof (Farrell and Maskin, 1989, Van Damme, 1989). Whether the withdrawal of cooperation that supports effort level \( x^{CE} \) is carried out via grim trigger strategies, repentance, or some combination of the two is irrelevant for our results—in particular, our results do not require pervasive community-wide punishments for individual deviations. The same comment will apply in later sections where cooperation is supported by the threat of both the withdrawal of cooperation and coercive punishment.

In practice, the most common way in which cooperation is withdrawn from deviators is probably *ostracism*—the exclusion of deviators from the benefits of cooperation, while the rest of the group continues to cooperate. Introducing ostracism into our model would have no effect on our results or their interpretation. In particular, suppose each player makes an additional choice \( \chi_i \) at the same time as her effort decision, which designates which other agents (if any) player \( i \) ostracizes and thus excludes from the benefits of her effort. (Alternatively, the whole group can ostracize individual \( k \) if \( \chi_i = \chi_j = k \) for all \( i, j \neq k \), i.e., if everyone agrees on whom to ostracize). In an efficient equilibrium, there is no ostracism on path, but deviators may be either permanently ostracized or ostracized until they repent by exerting effort without receiving any benefits as described in the previous paragraph. It is straightforward to verify that introducing ostracism in this way does not affect our equilibrium conditions, and hence does not affect any of our results, except that the
community can now discourage deviations with the threat of ostracism.\footnote{In a finite population, ostracizing one individual slightly reduces the maximum level of cooperation that can be sustained among the remaining players. This change does not affect equilibrium conditions or payoffs. For a discussion of various forms of ostracism in a model with imperfect private monitoring, see Ali and Miller (2016).}

### 3.2 Interpretation: Stateless Societies

Proposition 1, especially with the repentance or ostracism interpretation, provides a stylized representation of social order in stateless, small-scale societies. Equilibrium involves low levels of inequality across agents (in our simple model, no inequality at all). This is consistent with evidence from anthropology and archaeology on egalitarianism in most stateless societies (Bohannan and Bohannan, 1953, Boehm, 1999, Flannery and Marcus, 2014). Moreover, little coercion is used to support pro-social behavior (in the model, no coercion). Although there is continuous infighting, blood feuds, and endemic violence in many stateless societies (Chagnon, 1968, Boehm, 1986, LeBlanc and Register, 2003), there is limited use of coercion to support cooperation. Indeed, much violence in such societies appears to result from inter-group conflict (LeBlanc and Register, 2003), from various types of competition between males (Chagnon, 1968, Knauft, 1987, Marlowe, 2010), or from feuding between individuals or subclans that cannot be mediated in the absence of dispute resolution mechanisms (Boehm, 1986, Ember, 1978, Acemoglu and Robinson, 2019). Detailed ethnographic studies dating back to Radcliffe-Brown’s (1922) work on the Andamans in India do not find much evidence of coercive punishments to support cooperation in these societies (see, e.g., Briggs, 1970, on the Inuit, Woodburn, 1982, on the Hadza, or Wiessner, 2005, on the !Kung Bushmen; see Baumard, 2010, for a general discussion). In all of these cases, cooperation appears to be supported by a combination of low social regard directed at non-cooperators and the threat of withdrawal of future cooperation, for example via social isolation. The same appears to be true in societies with nascent but still-weak state institutions, such as Germanic tribes and subsequently Frankish states shortly after the fall of the Western Roman Empire, as well as early Anglo-Saxon England. In these cases, most infractions were punished by payments from perpetrators to victims or their families, for example via the wergeld as specified by the Salic Law of the Franks or King Alfred’s Law-Code (Drew, 1991, Acemoglu and Robinson, 2019). This arrangement closely resembles community enforcement supported by repentance and/or ostracism, as described above.\footnote{The example of wergeld raises the question of whether introducing monetary transfers would matter for the model. The answer is essentially no: so long as \( f_k(x) > 1 \) and \( f_k'(y) > 1 \) for effort levels that arise in equilibrium, it is more efficient to demand additional effort rather than on-path transfers, and replacing off-path “repentance effort” with “repentance transfers” would not affect any of our results.}

### 4 Elite Enforcement

We next turn to an analysis of the same environment, but now with a group of agents—the “elite”—that control the means of coercion in society.
4.1 Elite-Optimal Equilibria

Suppose now that $\alpha > 0$ (there are some elite agents in the population), $g \geq 0$ (coercive punishments are possible), and $\rho = 0$ (elite agents are themselves immune to coercion). In this game, the best equilibrium for normal agents and the best equilibrium for elite agents typically differ. As we are mainly interested in conditions under which elites themselves benefit from equality before the law, we focus for the moment on the best equilibrium for the elite.

**Proposition 2** Under elite enforcement,

1. Effort levels in every elite-optimal equilibrium are given by the solution to the problem

$$
\max_{x \geq 0, y \geq 0} (1 - \alpha) f_N(x) + \alpha f_E(y) - y \quad \text{s.t.}
$$

$$
x \leq \delta [(1 - \alpha) f_N(x) + \alpha f_E(y)] + g, \quad \text{(3)}
$$

$$
y \leq \delta [(1 - \alpha) f_N(x) + \alpha f_E(y)]. \quad \text{(4)}
$$

2. Constraint (3) binds at the optimum.

3. Let us denote the unique pair $(x, y) > (0, 0)$ such that both (3) and (4) bind by $(x^{EE}, y^{EE})$, and denote the solution to (2) subject to (3) and (4) by $(x^{EE}, y^{EE})$. Then we have

$$
\alpha f'_E(y^{EE}) + \delta (1 - \alpha) f'_N(x^{EE}) \leq 1 \text{ if } y^{EE} = 0, \quad \text{(5)}
$$

$$
\alpha f'_E(y^{EE}) + \delta (1 - \alpha) f'_N(x^{EE}) = 1 \text{ if } y^{EE} \in (0, \tilde{y}^{EE}), \quad \text{(6)}
$$

$$
\alpha f'_E(y^{EE}) + \delta (1 - \alpha) f'_N(x^{EE}) \geq 1 \text{ if } y^{EE} = \tilde{y}^{EE}. \quad \text{(7)}
$$

A number of features of this maximization problem are worth noting. First, the maximand, (2), is elite welfare, given by per-period benefits of cooperation $(1 - \alpha) f_N(x) + \alpha f_E(y)$ minus exerted effort $y$.

Second, (3) is the incentive constraint for a normal agent, while (4) is the incentive constraint for an elite agent. These constraints are intuitive: any player who deviates loses an expected benefit of $(1 - \alpha) f_N(x) + \alpha f_E(y)$ in the next period.\textsuperscript{15} Moreover, normal agents who deviate

\textsuperscript{14} If, as mentioned above, we interpret $y$ as the elite refraining from stealing and $f_E(y)$ as the damage that their extraction creates on normal agents, then (2) would need to be modified slightly by removing the $f_E(y)$ term from the objective function and the right-hand side of (4). This has no major impact on our main results.

\textsuperscript{15} This punishment can again be implemented by repentance, ostracism, or reversion to autarky by other players, and it applies to both normal and elite agents.

Among the different ways of withdrawing cooperation, repentance again has the advantage of being renegotiation-proof. In addition, in Acemoglu and Wolitzky (2019), we show that if punishments are costly to carry out, then another advantage of repentance is that it improves incentives for punishment. The implications of our analysis here are very different from that paper, for example because we here assume that punishments are costless and elites instead take productive actions.
face an additional coercive punishment of $g$. Normal agents are thus punished by both laws and norms. In contrast, there is no coercive punishment for elite agents, as elites are “above the law”, so this second term is not present in (4). However, as emphasized in footnote 3, elites are still constrained by social norms—if elites deviate from their prescribed behavior, the community withholds cooperation from them—which incentivize $y > 0$. Furthermore, in the best equilibrium for elites, normal agents are always required to work as hard as possible, so (3) binds.

Finally, (6) is the first-order condition with respect to $y$, once $x$ has been substituted out of the objective function using (3). This expression captures the fact that elites benefit in two ways from working harder. There is a direct marginal benefit of elites’ effort on other elites’ utility (the $\alpha f_E'(y)$ term). In addition, there is an indirect marginal benefit (the $\delta (1 - \alpha) f_N'(x)$ term): when elites work harder, future cooperation becomes more valuable, and thus normal agents are also incentivized to work harder (for fear of being excluded from the resulting increased benefits of cooperation). This indirect effect—and the complementarity between elite and normal agent effort it captures—works through the interplay of laws and norms, and is responsible for all of our comparative static results below. Since the indirect effect is due to forward-looking behavior by normal agents, it disappears when $\delta = 0$.

To better understand the indirect effect and to gain an intuition for the first-order condition for elite effort, note that each unit of marginal benefit created by the elites’ effort increases normal agents’ effort by $\delta$ units, which in turn provides $\delta (1 - \alpha) f_N'(x)$ units of benefit to both normal agents and elites. These units of benefit in turn increase normal agents’ effort by another $\delta^2 (1 - \alpha) f_N'(x)$ units, which provide $\delta^2 (1 - \alpha)^2 f_N''(x)$ units of benefit, and so on. The total marginal benefit to elites of increasing $y$ is thus given by the geometric series

$$\alpha f_E'(y) \left[ 1 + \delta (1 - \alpha) f_N'(x) + \delta^2 (1 - \alpha)^2 f_N''(x) + \ldots \right] = \frac{\alpha f_E'(y)}{1 - \delta (1 - \alpha) f_N'(x)}.$$ 

Equating this marginal benefit to the marginal cost of effort for the elite, which is 1, yields (6).\footnote{This role of coercive punishment $g$ in deterring deviations is somewhat similar to that in Acemoglu and Wolitzky (2011), where we assumed that employers/principals could use coercion in order to reduce the outside option of their employees/agents, thus forcing them to accept contracts that they would otherwise reject.}

\footnote{Another way of interpreting the cost to elites of increasing $y$ is that the resulting effort cost is borne only by elites, while the resulting benefits accrue to both elite and normal agents. This cost can be better understood by rewriting the first-order condition as

$$\alpha (f_E'(y) - 1) + \delta (1 - \alpha) f_N'(x) = 1 - \alpha,$$

where the $\alpha (f_E'(y) - 1)$ terms is the net direct benefit to the elite as a group from increasing all elites’ effort, $\delta (1 - \alpha) f_N'(x)$ is again the indirect benefit due to higher effort from normal agents, and $1 - \alpha$ is the share of benefits that are “wasted” on normal agents. This last term underscores the fact that the elites are unable to appropriate the full benefit of their increased effort because cooperation is a pure public good. However, this pure public good feature is not essential for our key qualitative results: in Section 8, we show similar results obtain when effort creates a mix of public benefits and private returns for one’s partner.}
4.2 Welfare

We next compare the welfare of normal and elite agents under community enforcement and elite enforcement. Let

\[ u_{CE} = f_N(x_{CE}) - x_{CE}, \]
\[ u_{EE}^N = (1 - \alpha) f_N(x_{EE}) + \alpha f_E(y_{EE}) - x_{EE}, \] and
\[ u_{EE}^E = (1 - \alpha) f_N(x_{EE}) + \alpha f_E(y_{EE}) - y_{EE}, \]
denote (elite-)optimal payoffs under community enforcement and elite enforcement, for (N)ormal and (E)lite agents. It is clear that elites prefer elite enforcement to community enforcement: \( u_{EE}^E \geq u_{CE} \). However, normal agents may or may not prefer elite enforcement to community enforcement.

The tradeoff is that under elite enforcement normal agents work harder than elites (i.e., \( x_{EE} > y_{EE} \)) and therefore receive a smaller share of the total social surplus \((1 - \alpha) (f_N(x) - x) + \alpha (f_E(y) - y)\) than under community enforcement, but total social surplus can be higher under elite enforcement than under community enforcement because the threat of coercive punishment increases the maximum sustainable effort level (i.e., \( \min \{x_{EE}, y_{EE}\} > x_{CE}\)).

We next ask how the comparison between \( u_{CE} \) and \( u_{EE}^N \) depends on \( g \). This dependence is ambiguous in general, but it can be characterized when the fraction of elite agents, \( \alpha \), is small.

**Proposition 3** Assume \( f'_E(0) < \infty \). There exists \( \bar{\alpha} > 0 \) such that if \( \alpha < \bar{\alpha} \), then there exists \( g^* \) (possibly equal to 0) such that \( u_{EE}^N > u_{CE} \) if \( g < g^* \) and \( u_{EE}^N < u_{CE} \) if \( g > g^* \).

The proof shows that when \( \alpha \) is sufficiently small that elites do not find it in their interest to exert effort under elite enforcement, normal agents prefer elite enforcement to community enforcement if and only if coercive capacity \( g \) is below a threshold \( g^* \). The logic is that a small but positive level of \( g \) can increase normal-agent effort from the inefficiently low level arising under community enforcement towards the first-best level, but that when \( g \) is too high normal agents will be forced to exert very high (above first-best) effort levels and are thus worse off than under community enforcement. Thus, while the output-increasing effect of state enforcement can dominate if coercion of limited, when the extent of coercion is very high (for example, as in ancient empires relying on large-scale labor coercion, such as Egypt or Sparta) the inequality effect dominates and normal agents are worse off.

Finally, it is useful to note that the elite enforcement model includes the special case \( g = 0 \) where there is no coercive technology but elites and normal agents may still be treated asymmetrically. In other words, if \( \alpha > 0 \) while \( g = 0 \), the model allows political hierarchy—and potentially some

\(^{18}\)To see that \( y_{EE} > x_{CE} \), note that \( x_{CE} \) is the positive root of the concave function \( \delta(f(x) - x) \), while \( y_{EE} \) is the positive root of the concave function \( \delta[(1 - \alpha) f(y + \alpha g) + \alpha f(y)] - y \). The latter function is everywhere greater than the former, so it has the greater root.
degree of inequality—even in the absence of coercion. Note, however, that when $g = 0$, we have $\tilde{x}_{EE} = \tilde{y}_{EE}$; that is, if it is optimal for elites to work at the maximum sustainable level when $g = 0$, then egalitarianism is their most preferred option. While it is not always optimal for elites to exert maximal effort, we note that $\tilde{x}_{EE}$ and $\tilde{y}_{EE}$ are increasing in $g$, so (7) is easiest to satisfy—and thus equal levels of effort are most likely—when $g$ is small. This result, that egalitarianism is most likely to arise when $g$ is small, establishes our earlier claim that low coercion and equality go hand-in-hand: when the elite cannot use coercion effectively, it is optimal from their viewpoint for them to exert similar effort to normal agents. This is also the reason why we believe the low coercion, high inequality cell in Table 1 is not well-populated.

4.3 Interpretation

In contrast to community enforcement, elite enforcement involves high inequality and high coercion. Both of these are characteristic of early societies that developed state institutions (Johnson and Earle, 2000, Flannery and Marcus, 2014). These features are also the hallmarks of what North, Wallis and Weingast (2009) call limited access orders, where a well-defined elite monopolizes the means of violence and enjoys rents, as well as of extractive economic institutions (Acemoglu and Robinson, 2012), which empower elites to enjoy unfair advantages in economic relations.

Note additionally that unequal treatment before the laws generates unequal norms. The threat of coercive legal punishment induces normal agents to exert higher effort, and normal agents who fail to deliver this higher level of effort are punished by both coercion and the withdrawal of cooperation.

The feudal social order, which was widespread throughout medieval Europe, provides one clear illustration of an elite-dominated system. A defining feature of feudal society was a clear distinction between the military elite, which dominated the means of coercion, and the rest of society. Feudal society was highly hierarchical, and the social hierarchy was supported by access to the means of coercion, control of land, and custom, all privileging the elite (lords, knights, or nobles). At the bottom of the hierarchy were the “serfs”, who over the course of the medieval era evolved into hereditary bonded laborers without freedom of movement or occupational choice. In the words of the historian Richard Southern (1952, p. 98), “There were two great and universal divisions in the society…: between free men and serfs, and between free men and noblemen.”

The laws that applied to lords were different than those that applies to other freemen and to unfree serfs. Though coercion played a critical role in the highly militarized feudal society, custom and norms were also crucial (as in our model) and imposed duties on the elite as well as the citizens. These norms, however, were unequal and privileged the elite in many domains (Bloch, 1939, Part VI).

---

19The famous historian of this system, Marc Bloch (1939, Part IV, pp. 153-155), emphasizes the pervasive pattern of dependence and vassalage in feudal society and writes, “the subordinate was often simply called the ‘man’ of [the] lord; or sometimes more precisely, his ‘man of mouth and hands’.”
Our framework sheds light on an important debate concerning whether the transition from stateless societies to societies with more organized institutions and coercion was welfare-improving for the population at large because it encouraged better cooperation and dispute resolution (as maintained by various social contract theories going back to Thomas Hobbes and John Locke; see also Huntington, 1968, Bates, 2001, Fukuyama, 2011), or welfare-reducing for most because it led to exploitation by the elite (as maintained by Scott, 2017, and suggested by evidence of affluence and relatively good health among some stateless societies, e.g., Sahlins, 1974, Suzman, 2017). Our analysis shows either outcome is possible. Under elite enforcement (relative to community enforcement), there is greater inequality favoring the elite, which tends to make normal agents worse off. At the same time, because higher effort benefits everyone, normal agents may become better off as well. Elite agents are always better off under elite enforcement, because they benefit both from the higher effort of normal agents and from a privileged position resulting from their monopoly on coercion and their above-the-law status. This feature is also consistent with the existing archaeological and historical evidence (e.g., Flannery and Marcus, 2014).

These opposing forces were apparent in feudal society. Many historians trace the origin of feudalism to the need for protection and stability in an age when Europe was beset by warfare and suffered continual attacks from nomadic Hungarian and Muslim warriors. There is no doubt that the feudal order provided some degree of security in this environment. Nevertheless, the evidence is also clear that serfs were heavily exploited, and whenever they could (for example, after the population collapse following the Black Death), they tried to reassert their freedom and break away from their servile obligations (e.g., North and Thomas, 1973, Benedictow, 2004).

5 Equality Before the Law

We now turn to our main focus: how equality before the law emerges endogenously. We thus now model the extent to which elites are subject to coercive punishments as a choice variable.

5.1 Elite-Optimal Equilibria

We suppose that $\rho \in [0,1]$ is a choice variable for the elite, and continue to focus on the elite-optimal equilibrium. The interpretation is that the elite hold the political power to choose both the institutional environment ($\rho$) and the equilibrium. We assume that a cost $\varepsilon \rho$ is incurred by all agents when the extent of equality before the law is $\rho$. This may represent the cost of establishing the institutional foundations for greater equality before the law. In what follows, we take $\varepsilon \to 0$.

---

20 In doing so, we are also implicitly characterizing the best equilibrium for the elite for any fixed value of $\rho$. 

15
which simplifies the analysis.\textsuperscript{21} Under this assumption, the problem for the elites becomes

\[
\max_{x \geq 0, y \geq 0, \rho \in [0,1]} \quad (1 - \alpha) f_N(x) + \alpha f_E(y) - y \quad \text{s.t.}
\]
\[
x \leq \delta [(1 - \alpha) f_N(x) + \alpha f_E(y)] + g,
\]
\[
y \leq \delta [(1 - \alpha) f_N(x) + \alpha f_E(y)] + \rho g,
\]

where (10) is the incentive compatibility constraint for elites, which must hold with equality if \( \rho > 0 \). Here (9) is identical to (3), while (10) differs from these constraints only in that an elite agent’s minmax payoff is \(-\rho g\) rather than \(-g\).

Let us denote the unique solution to the elites’ problem—corresponding to the optimal equilibrium under endogenous equality before the law with minimal \( \rho \)—by \((x^{EL}, y^{EL}, \rho^*)\). Here uniqueness follows from concavity, and the superscript \( EL \) stands for “Equality before the Law”.

To characterize the solution, first note that it is always optimal for (9) to bind, as increasing \( x \) increases the objective and relaxes (10). Hence, \( x^{EL} = x^* (y^{EL}) \), where again \( x^* (y) \) is the value of \( x \) that makes (9) hold with equality. Let \((\bar{x}^{EL}, \bar{y}^{EL})\) be the unique pair \((x, y)\) such that (10) binds with \( \rho = 1 \). That is, \((\bar{x}^{EL}, \bar{y}^{EL})\) are the greatest sustainable effort levels under equality before the law. Note that \( \bar{x}^{EL} = \bar{y}^{EL} \), which implies that the maximum sustainable effort level for normal and elite agents is the same under full equality before the law.

The following proposition provides our main result. It characterizes the elite-optimal level of equality before the law and the resulting equilibrium effort levels.

**Proposition 4** Every elite-optimal equilibrium takes one of the following three forms:

1. **Elite enforcement:** \( \rho^* = 0 \), \((x^{EL}, y^{EL}) = (x^{EE}, y^{EE})\), and

\[
\alpha f'_E(y^{EE}) + \delta (1 - \alpha) f'_N(x^{EE}) \leq 1.
\]

2. **Partial equality before the law:** \( \rho^* \in (0, 1) \), \( y^{EL} \in (\bar{y}^{EE}, \bar{y}^{EL}) \), \( x^{EL} = x^* (y^{EL}) \in (\bar{x}^{EE}, \bar{x}^{EL}) \), (10) binds, and

\[
\alpha f'_E(y^{EL}) + \delta (1 - \alpha) f'_N(x^{EL}) = 1. \tag{11}
\]

\textsuperscript{21}The presence of this cost removes an uninteresting multiplicity. Since increasing \( \rho \) relaxes the incentive constraint of the elite and we focus on the elite-optimal equilibrium, the elite would be willing to choose \( \rho = 1 \) (full equality before the law) and not punish themselves: intuitively, the elite are happy to allow themselves to be subject to coercion, provided the equilibrium specifies they are never actually coerced. This is a consequence of the fact that the elite are never punished along the equilibrium path. The small cost of increasing \( \rho \) rules out this rather artificial possibility and implies that the elites always choose the smallest level of \( \rho \) when indifferent. One could also obtain the same conclusion by introducing a small probability that the elites suffer punishment on path and taking this probability to zero.
3. Full equality before the law: \( \rho^* = 1 \), \( (x^{EL}, y^{EL}) = (\bar{x}^{EL}, \bar{y}^{EL}) \) (in particular, \( x^{EL} = y^{EL} \)), and

\[
\alpha f'_E (\bar{y}^{EL}) + \delta (1 - \alpha) f'_N (\bar{x}^{EL}) \geq 1.
\]

The maximization problem (8) differs from (2) only in that \( \rho \) is now a choice variable, rather than being fixed exogenously at 0. As in the earlier problem, the incentive compatibility constraint of normal agents, (9), always binds, and that of elite agents, (10), binds only if the best equilibrium for the elite involves maximum elite agent effort. Hence, if (10) with \( \rho = 0 \)—or equivalently the corresponding constraint (4) under elite enforcement—is slack, then elites have no interest in committing themselves to a higher level of effort, and instead prefer to remain in the elite enforcement regime with \( \rho = 0 \). In contrast, if (4) binds under elite enforcement (or equivalently, if (7) holds with strict inequality), then the elites opt for at least partial equality before the law, where the optimal level of equality before the law is just sufficient to commit themselves to the effort level \( y^{EL} \) satisfying the first-order condition (11).  

Finally, in the case where \( \alpha f'_E (\bar{y}^{EL}) + \delta (1 - \alpha) f'_N (\bar{x}^{EL}) \geq 1 \), elites prefer full equality before the law. In this case, the best equilibrium from the viewpoint of the elites involves \( x = y \); that is, we obtain not only equality before the law but also equal norms, and indeed completely equal allocations. This gives another way of viewing the last part of the proposition: the elite prefer to establish full equality before the law only when they are willing to work as hard as normal agents.

We can also provide a diagrammatic representation and intuition for Proposition 4. Recall first that \( \rho^* \) is either 0 or the value of \( \rho \) that binds (10). We can thus omit (10) and rewrite the elites’ problem, (8), as

\[
\max_{x \geq 0, y \in [0, \bar{y}]} (1 - \alpha) f_N (x) + \alpha f_E (y) - y
\]

subject to (9), where \( \bar{y} = \bar{y}^{EE} \) under elite enforcement and \( \bar{y} = \bar{y}^{EL} \) under endogenous choice of \( \rho \). We illustrate this problem in Figure 1. The thick curve represents combinations of normal agent and elite effort that satisfy the normal agents’ incentive compatibility constraint, (9), as an equality. This curve intersects the 45° line at the point \( (\bar{x}^{EL}, \bar{y}^{EL}) \), which corresponds to fully equality before the law, \( \rho^* = 1 \) (and equal effort from normal and elite agents). The point \( (x^{EE}, y^{EE}) \), corresponding to elite enforcement with \( \rho^* = 0 \), is plotted as well. The figure also superimposes the indifference curves of (12), which are convex (since (12) is concave). The point of tangency, if

\[\text{[22]}\]

The intuition for this first-order condition with endogenous \( \rho \) is the same as for the one with \( \rho = 0 \) given in (6): the direct marginal benefit to elites of increasing their effort is \( \alpha f'_E (y) \), and the indirect marginal benefit—coming through the induced increase in the maximum incentive compatible level of normal agent effort—is \( \delta (1 - \alpha) f'_N (x) \). The first-order condition sets the total marginal benefit of \( \alpha f'_E (y) + \delta (1 - \alpha) f'_N (x) \) equal to the total marginal cost of 1.

\[\text{[23]}\]

Normal and elite agents exert the same effort even though \( f_N \) and \( f_E \) may differ. This is because effort levels of the two types of agents under equality before the law are determined by their binding incentive compatibility constraints, which are identical and thus imply the same level of effort. This is no longer the case in Section 8, where elites may receive greater benefits from cooperation.
any, between these indifference curves and the boundary of (9) gives the combination of \((x, y)\) that is optimal from the viewpoint of the elite; such a point of tangency corresponds to an intermediate value of \(\rho^* \in (0, 1)\). When there is no tangency, the highest indifference curve is reached either at the corner where \((x, y) = (x^{EL}, y^{EL})\) with full equality before the law \((\rho = 1)\), or at the point where \((x, y) = (x^{EE}, y^{EE})\) with elite enforcement \((\rho = 0)\).

5.2 Welfare

We next consider the implications of equality before the law for the welfare of normal and elite agents. Let \(u_N^{EL}\) and \(u_E^{EL}\) be normal and elite agents’ utility under the endogenous (elite-optimal) choice of equality before the law. Clearly, \(u_E^{EL} \geq u_E^{EE}\), with strict equality if \(\rho^* > 0\): this follows because elites have an extra choice variable under endogenous equality before the law. More interestingly, we have:

**Proposition 5** \(u_N^{EL} \geq u_N^{EE}\), with strict equality if \(\rho^* > 0\). In addition, if \(\rho^* = 1\) then \(u_N^{EL} > u_N^{CE}\).

That is, under the elite-optimal equilibrium with endogenous equality before the law, normal agents are always better-off than under elite enforcement. This follows because inequality is reduced and effort among all individuals is increased. When full equality before the law is optimal for elites, normal agents are also better-off than under community enforcement.
5.3 Interpretation

Like elite enforcement, equality before the law uses threats of coercive punishment to encourage pro-social behavior. However, in contrast to elite enforcement, it involves a low degree of inequality: elite agents are not treated in a privileged manner. In this respect, our model captures an ideal aspired to by most modern Western constitutions, which simultaneously enshrine equal treatment of all individuals and strong legal enforcement. As emphasized by Hart (1961), legal enforcement critically depends on—and simultaneously shapes—society’s norms. This interdependence is captured in our model by the synergy between coercive punishments and repeated game incentives.

Our model also resonates with the emphasis on rule of law in the institutional economics literature. Equality before the law in our model has much in common with the ideal of rule of law of Hayek (1960), who emphasized the defining role of equal application of laws and equal protection from coercion. It is a critical component of open access orders as described by North, Wallis and Weingast (2009), where society is governed according to the rule of law, and access to the means of violence is separated from access to rents. And it is also a key aspect of inclusive economic institutions in Acemoglu and Robinson (2012), which depend on a level economic playing field among all individuals and thus the removal of various privileges before the law. Indeed, the evolution of many Western societies towards more democratic and inclusive institutions can be viewed precisely as such a process of stripping away the privileges of elites.

The model also implies that, as exogenous parameters change such that the optimal equilibrium for elites transitions from elite enforcement to full equality before the law, it typically passes through a substantial region of partial equality before the law \( \rho \in (0, 1) \). This is consistent with the historical cases we discuss in Section 7 such as the Meiji Constitution in Japan and Tanzimat reforms in the Ottoman Empire, which took important steps towards equality before the law, but certainly did not bring full equality before the law to Ottoman and Japanese societies.

Finally, we have so far assumed that if the best equilibrium for the elite involves some degree of equality before the law—\( \rho > 0 \)—then the elite can freely choose and commit to such an arrangement. An important question is how this can be secured in practice: that is, how can the coercion and enforcement roles in society be separated from elite status? A vital aspect of the solution is to transfer the means of coercion from elites to agents specialized in law enforcement, similar to what the Meiji government did by disarming the samurai and creating a professional police force. A more modern version of the same solution is to create a (sufficiently independent) government bureaucracy and judiciary to resolve conflicts and decide whom should be subject to punishment. In both cases, the practical challenge is to ensure the independence and impartiality of the agents charged with law enforcement or judiciary functions. We leave a more in-depth investigation of this important issue for future work.
Figure 2: The indifference curves of the elite become flatter as we move to the right along a horizontal line. An increase in $g$ shifts out the red curve representing the boundary of the incentive compatibility constraint (9) to the right, and thus leads to a new point of tangency with greater $x$ and lower $y$, and thus lower $\rho^\ast$.

6 Comparative Statics: Towards Equality Before the Law

We now turn to comparative statics on how the elite-optimal levels of production and equality before the law vary with parameters. We illustrate our most important comparative static results with several historical examples in the next section.

6.1 Comparative Statics for Coercive Capacity

Our most important comparative static says that an increase in coercion increases economic inequality and decreases equality before the law.

**Proposition 6** An increase in coercive capacity $g$ leads to an increase in normal agent effort, a decrease in elite agent effort, and a decrease in equality before the law.

Formally, $x^{EE}$ and $x^{EL}$ are strictly increasing in $g$, $y^{EE}$ and $y^{EL}$ are nonincreasing in $g$, and $\rho^\ast$ is nonincreasing in $g$. In addition, if $\delta > 0$ and the solutions are interior, then the comparative statics on $y^{EE}$, $y^{EL}$, and $\rho^\ast$ are strict.

Figure 2 provides a diagrammatic intuition for Proposition 6. An increase in $g$ has no impact on the indifference curves of the elite, but shifts the boundary of (9) to the right. The indifference
curves of the elite become less steep as we move to the right along a horizontal line. Consequently, shifting (9) leads to a new tangency point with not only greater $x$, but also lower $y$. Lower elite effort then translates into a lower level of equality before the law.

A complementary intuition is that coercive punishments and incentives provided by norms/threat of withdrawal of cooperation are substitutes at the margin. The greater is $g$, the less need there is for additional incentives coming from norms, and this allows the elite reduce $y$. More precisely, recall that part of the elites’ incentive to choose greater effort $y$ is that this indirectly increase normal agent effort $x$. An increase in $g$ raises $x$ for a fixed level of $y$. Because $f_N$ is concave (i.e., there are diminishing returns to effort), the term $\delta (1 - \alpha) f'_N(x)$ in (6), which captures this indirect effect, declines when $x$ increases. This encourages the elite to choose a lower $y$. Since increasing $\rho$ is a way to raise $y$ (by making elites subject to greater coercive punishments), an increase in $g$ also leads to a reduction in $\rho$. The fact that this comparative static ceases to be strict when $\delta = 0$ confirms this intuition, since in this case there are no repeated game considerations and hence no indirect effect.

This comparative static, which is one of our main results, will be illustrated in the next section with the rise of equality before the law in ancient Athens and in Britain in the 19th century.

6.2 Comparative Statics for Political Power

Our previous comparative static focused on restrictions on the extent of coercion available to elites, while maintaining elites’ political power. But many of the same social changes that restrict elites’ access to coercion—most notably the emergence of mass democratic politics—also reallocate political power towards normal agents (see the discussion and references in Acemoglu and Robinson, 2006). We now show that a decline in the relative political power of the elite also contributes to the emergence of equality before the law. We establish this result in the simplest possible fashion (without introducing a micro-founded model of the political power of the elite) by simply considering the set of equilibria that maximize a weighted average of the utilities of the elite and normal agents, and then reducing the weight of the elite in this social welfare function. In the process, we also confirm that none of our results so far depend on focusing on the best equilibrium for the elite.

Our first result establishes that under elite enforcement (more generally, for any fixed level of equality before the law $\rho$), a more equal distribution of political power typically leads to higher effort for both normal agents and elites, and hence higher output. In particular, this holds whenever normal agents’ incentive constraints bind (e.g., whenever effort is below the first-best level). The

---

24 This follows because the slope of the indifference curve is $-\frac{f'_N(x)}{f'_E(y)}$, which gets flatter as $x$ increases (since $f_N$ is concave).

25 The direct, positive effect of an increase in $g$ on $x$ always outweighs the indirect, negative effect coming through the decrease in $y$, so $x$ is indeed increasing in $g$.

26 There is an exception to this: it is possible that $dp^*/dg$ is strictly negative even when $\delta = 0$, as the value of $\rho$ that binds (3) is decreasing in $g$ even when $\delta = 0$. 

---
intuition is that elite agents work more at the optimum under more equal Pareto weights, and this in turn induces higher effort from normal agents. Thus, unequal political power reduces production.

**Proposition 7** Under elite enforcement, let \((x^{EE}(\gamma), y^{EE}(\gamma))\) denote the optimal equilibrium effort levels with Pareto weight \(\gamma\) on the elite, given by the solution to

\[
\max_{x \geq 0, y \geq 0} (1 - \alpha) f_N(x) + \alpha f_E(y) - (1 - \gamma) x - \gamma y
\]

subject to (3) and (4). For all Pareto weights \(\gamma > \gamma' \geq \alpha\), if \(x^{EE}(\gamma) < x^{FB}\) then \(x^{EE}(\gamma) \leq x^{EE}(\gamma')\) and \(y^{EE}(\gamma) \leq y^{EE}(\gamma')\).

Note that the assumption \(\gamma, \gamma' \geq \alpha\) says that the Pareto weights favor the elite.

In terms of Figure 1, an increase in the Pareto weight of the elite has no impact on the constraint set and rotates the indifference curves clockwise, thus shifting the equilibrium to a point with lower \(x\) and \(y\) along (9). The resulting decline in elite effort—combined with an increase in elite utility, which makes the carrot of future cooperation more effective for the elite and thus reduces the need for the elite to face coercive punishment—then leads to a reduction in equality before the law.

**Proposition 8** Let \((x^{EL}(\gamma), y^{EL}(\gamma), \rho^*(\gamma))\) denote the optimal equilibrium levels of effort and equality before the law with Pareto weight \(\gamma\) on the elite, given by the solution to

\[
\max_{x \geq 0, y \geq 0, \rho \in [0, 1]} (1 - \alpha) f_N(x) + \alpha f_E(y) - (1 - \gamma) x - \gamma y
\]

subject to (9) and (10). For all Pareto weights \(\gamma > \gamma' \geq \alpha\), if \(x^{EL}(\gamma) < x^{FB}\), then \(x^{EL}(\gamma) \leq x^{EL}(\gamma')\), \(y^{EL}(\gamma) \leq y^{EL}(\gamma')\), and \(\rho^*(\gamma) \leq \rho^*(\gamma')\).

### 6.3 Comparative Statics for the Returns to Effort

Our next comparative static analyzes how changes in the nature of the production function affect the transition to equality before the law. As discussed in the Introduction, several historical examples—most notably the episodes of “defensive modernization” in 19th-century Prussia, Japan, and the Ottoman Empire—suggest that reforms leading to greater equality before the law take place when a society is faced with external threats that necessitate intensification of industrialization or armament. In terms of our model, this corresponds to an increase in the slope of the functions \(f_N\) and \(f_E\), that is, an increase in marginal returns to effort (the need to increase production), but not average returns (the economy’s productivity).

The distinction between marginal and average returns is important for this comparative static, because increasing marginal returns encourages greater effort from both normal and elite agents (inducing greater equality before the law in equilibrium), while increasing average returns makes
retaining their privileged position more attractive for the elite. In this subsection, we therefore focus on rotations of the \( f_N \) and \( f_E \) functions that isolate the first effect, and show that such changes in the benefit production functions lead to greater equality before the law.

Suppose the production functions \( f_N \) and \( f_E \) are parameterized by \( \theta \in [0, 1] \). Let \((x_0, y_0, \rho_0)\) denote the elite-optimal equilibrium given \( \theta_0 \in (0, 1) \), and let \((x^* (\theta), y^* (\theta), \rho^* (\theta))\) denote the elite-optimal equilibrium as a function of \( \theta \). Assume \( f_N \) and \( f_E \) are twice continuously differentiable in \((x, \theta)\).

**Proposition 9** Suppose that increasing \( \theta \) raises marginal returns to effort at \( x_0 \) and \( y_0 \) while decreasing average returns to effort at \( x_0 \) and \( y_0 \): that is,

\[
\frac{\partial^2}{\partial x \partial \theta} f_N (x_0, \theta_0) \geq 0, \quad \frac{\partial^2}{\partial y \partial \theta} f_E (y_0, \theta_0) \geq 0, \quad \frac{\partial}{\partial \theta} f_N (x_0, \theta_0) \leq 0, \quad \frac{\partial}{\partial \theta} f_E (y_0, \theta_0) \leq 0.
\]

Assume \( y^* (\theta) \) and \( \rho^* (\theta) \) are differentiable in \( \theta \) at \( \theta = \theta_0 \). Then these derivatives are both non-negative: that is, as marginal returns to effort increase, elite agents exert more effort, and equality before the law increases.

The comparative statics on \( x^* \) is ambiguous, because the positive incentive effect of an increase in \( y^* \) is offset by the negative incentive effect of a reduction in average returns for fixed \( x^* \) and \( y^* \).

The result that \( \frac{d \rho^*}{d \theta} \) is non-negative is somewhat subtle. Suppose increasing \( \theta \) raises marginal returns while leaving average returns unchanged (a case allowed by the proposition). It is quite intuitive that this leads to an increase in \( x^* \) and \( y^* \). But why does this encourage greater equality before the law? In other words, why is the increased carrot of future cooperation not enough to justify the resulting higher level of elite effort? Intuitively, increasing \( \theta \) raises both the level of elite effort collectively preferred by the elite group \((y^*)\) and the level of effort that each elite agent finds it individually optimal to exert. But the latter increase will always fall short of the former, because it is incentivized only by the increased benefits that elite agents enjoy in equilibrium, and since the initial allocation was chosen to maximize net benefits to the elite, the implied increase in elite effort from these greater benefits will be small. Hence to achieve the desired increase in \( y^* \), the elite collectively need to make themselves subject to greater coercive punishments.\(^{27}\)

\(^{27}\)To see this in a little more detail, denote total benefits from cooperation (gross of costs) by \( B = (1 - \alpha) f_N (x^* (\theta), \theta) + \alpha f_E (y^* (\theta), \theta) \). Since (10) binds at \( \rho^* \), we have

\[
g \frac{d \rho^*}{d \theta} = \frac{dg}{d \theta} - \delta \frac{dB}{d \theta}.
\]

At the elite-optimal equilibrium, we have \( \frac{\partial B}{\partial \theta} = 1 \). Thus

\[
\frac{dB}{d \theta} = \frac{\partial B}{\partial \theta} + \frac{\partial B}{\partial y^*} \frac{dy^*}{d \theta} \leq \frac{\partial B}{\partial y^*} \frac{dy^*}{d \theta} = \frac{dy^*}{d \theta}.
\]

Hence, \( g \frac{d \rho^*}{d \theta} \geq (1 - \delta) \frac{dy^*}{d \theta} \). As \( \frac{dy^*}{d \theta} \geq 0 \) and \( \delta < 1 \), this implies \( \frac{d \rho^*}{d \theta} \geq 0 \). The proof of the proposition spells this argument out in greater detail.
Overall, the substantive conclusion of this subsection is that an increase in the marginal returns to effort, which may result from a change in technology or a situation of national emergency, encourages greater equality before the law. This comparative static is another one of our major results and will be discussed in detail in the next section.

### 6.4 Comparative Statics for Inequality

In this subsection, we slightly modify our baseline setup to discuss the effects of economic inequality between elite and normal agents on the emergence of equality before the law. Recent increases in wealth and income inequality around the world (e.g., Atkinson, Piketty, and Saez, 2011) have raised concerns about whether a system based on equal opportunity—and in our setting, equality before the law—can survive in a highly unequal society. Scheidel (2017) argues this has not been possible historically, and only war and revolution have tended to limit inequality. There are indeed several historical cases where early steps towards equality before the law have been reversed following increases in economic and political inequality, for example, in the Roman Republic and medieval Venice (e.g., Acemoglu and Robinson, 2012, and Puga and Treﬁer, 2014). We now show that one type of increase in inequality—where the elite get richer while normal agents do not—makes equality before the law less likely to emerge (and perhaps harder to maintain) in our model.

For this exercise, we return to the random matching version of our model in Section 2.2 where each agent’s effort generates benefits for their partner and effort decisions are made under anonymity. We then modify this setup by introducing heterogeneous endowments for elite and normal agents, and then investigate the implications of an increase in the endowment of the elite holding those of normal agents constant (and other combinations).

Let us now interpret effort \( x_i \) by agent \( i \) as producing \( f_i(x_i) \) units of a non-storable consumption good for her partner (where \( f_i = f_N \) or \( f_E \) depending on one’s type). In addition, each agent has a per-period endowment of consumption goods, which equals \( e_N \) for normal agents and \( e_E \) for elites. Agents have utility function over consumption \( u_i(\cdot) \) satisfying \( u_i' > 0, u_i'' < 0 \) (where again \( u_i = u_N \) or \( u_E \) depending on the agent’s type). Consequently, if agent \( i \) has endowment \( e_i \) and exerts effort \( x_i \) while her partner exerts effort \( x_j \), agent \( i \)’s payoff is

\[
u_i(e_i + f(x_j)) - x_i.
\]

The next result shows that an increase in elites’ endowments decreases production. The intuition is that increasing \( e_E \) decreases elites’ marginal utility of consumption, thus reducing both the direct and indirect benefits of increasing \( y \).

**Proposition 10** An increase in elites’ endowments \( e_E \) leads to lower normal and elite agent effort. Formally, \( x^{EE}, x^{EL}, y^{EE}, \) and \( y^{EL} \) are nonincreasing in \( e_E \).
The modified problem here again generates a set of convex indifference curves in Figure 1, and an increase in elites’ endowments has no effect on the constraint set but rotates the indifference curves clockwise, decreasing both elite and normal agent effort, and consequently reducing equality before the law.\(^{28}\)

To see the corresponding effect on equality before the law, recall that \(\rho\) is defined so that the elite incentive compatibility constraint, now given by

\[
y \leq \delta \left[(1 - \alpha) u_E (e_E + f_N (x)) + \alpha u_E (e_E + f_E (y)) - u(e_E)\right] + \rho g,
\]

binds. An increase in \(e_E\), which from Proposition 10 reduces \(y\), creates two opposing effects on this constraint. On the one hand, via the first two terms on the right-hand side, it relaxes the constraint and thus pushes for a lower value of \(\rho\). On the other hand, via the \(-u(e_E)\) term, it tightens the constraint. This offsetting effect comes from the fact that a higher endowment for the elites improves their payoffs under autarky, making a deviation more tempting. Greater equality before the law may now be useful to counteract this heightened temptation to deviate. However, if we interpret the allocation of endowments as being socially determined as well—so that deviators can be ostracized and excluded from having access to or enjoying the benefits from their endowments—then this second effect disappears. In this case, greater elite endowments (and greater inequality) unambiguously reduce equality before the law.

### 6.5 Comparative Statics for the Size of the Elite

Our next comparative static says that a larger elite prefers a higher level of equality before the law, i.e., higher \(\rho\). This is consistent with the argument of North, Wallis and Weingast (2009) that first establishing some level of equality before the law among a larger segment of the elite (which we interpret here as increasing the size of the elite) is a key doorstep condition for subsequently extending equality before the law to the broader population.

**Proposition 11** Assume \(f_N = f_E = f\). Then an increase in the size of the elite, \(\alpha\), leads to an increase in elite agent effort and an increase in equality before the law. Formally, \(y^{EE}\), \(y^{EL}\), and \(\rho^*\) are nondecreasing in \(\alpha\). If the solutions are interior, then the comparative statics are strict.

To see the intuition for this result, note that an increase in \(\alpha\) reduces \(x\) for a fixed level of \(y\), while also raising the marginal benefit to elites of higher \(y\) for a fixed level of \(x\) and \(y\). As \(f\) is

---

\(^{28}\) Proposition 10 considers increasing elites’ endowments while holding fixed the endowment of normal agents. If instead \(e_N\) declines as \(e_E\) increases, the comparative static becomes ambiguous: With a lower endowment, normal agents work harder and the greater returns that this creates for the elite discourages them from exerting effort, reinforcing the result in Proposition 10. However, with a lower endowment, the sensitivity of normal agents’ effort to elite effort increases and this might encourage the elites to increase their effort. Nevertheless, it can be shown that if \(u_N\) is not very concave, this second effect is dominated and thus the conclusion of Proposition 10 continues to apply.
concave, the net effect is to raise the marginal benefit to elites of increasing \( y \). The comparative static with respect to \( \alpha \) is strict even if \( \delta = 0 \), as changing \( \alpha \) influences the direct effect term \( \alpha f' (y^{EL}) \) in (6) in addition to the indirect effect. Finally, note that the overall effect of an increase in \( \alpha \) on \( x \) is ambiguous, because the direct, negative effect on \( x \) may be offset by the indirect, positive effect coming through the increase in \( y \).

7 Reinterpreting the Rise of Equality before the Law

In this section, we discuss several historical examples that both illustrate our key comparative static results and can be reinterpreted in light of our results. Our focus will be on the two key comparative static results presented in the previous section—those with respect to limits on coercive punishments and changes in returns to effort.

7.1 Equality before the Law in Ancient Athens

The beginning of the rise of equality before the law in Athens can be dated to the 6th century BC, in particular to the appointment of Solon to the chief executive position, *Archon*, in 594 BC. Solon was brought to power during a period of significant discord between elite families and regular Athenian citizens, and his charge was to restructure Athenian institutions to provide greater protection to citizens. Though Solon’s remit was revolutionary, he was acceptable to the elites partly because the need for institutional changes was widely felt that at the time. Aristotle, in his *The Constitution of Athens*, quotes Solon as stating: “To the people I gave as much privilege as was sufficient for them, neither reducing nor exceeding what was their due. Those who had power and were enviable for their wealth I took good care not to injure. I stood with my shield outstretched, and both were safe in its sight. And I would not that either should triumph, when the triumph was not with the right” (2009, p. 14).

Critical among Solon’s reforms were several laws protecting citizens against various abuses. First, he eliminated debt peonage. Second, he made enserfing an Athenian citizen illegal. Third, he implemented a fairly radical land reform. Osborne (2009, p. 211) interprets this as Solon “…freeing the tenants from landowners, giving them the land they owned, and turning Attica into the land of small farmers”. Fourth, he implemented various judicial reforms making it easier for Athenians to access courts and seek justice. Fifth, he enacted a hubris law, which made it illegal for people to act hubristically towards other Athenians, including slaves (Ober, 2015, pp. 150-152). Prior to this reform, it appears that it was commonplace for elites to humiliate and intimidate regular citizens, and reining in such behaviors was an important step towards equality before the law.
Finally, Solon also implemented a series of reforms that made Athens’ political institutions more democratic and attempted to increase the capacity of the Athenian state. Aristotle describes Solon’s most important reforms as follows, emphasizing the importance of access to justice and a form of equality before the law:

“There are three points in the constitution of Solon which appear to be the most democratic features: first and most important, the prohibition of loans on the security of the debtor’s person; secondly, the right of every person who so willed to claim redress on behalf of anyone to whom wrong was being done; thirdly, the institution of the appeal to the jurycourts; and it is to this last, they say, that the masses have hold their strength most of all” (2009, p. 12).

Ober summarizes this as follows:

“Powerful officials thus became the equals of ordinary citizens before the law, a development with profound implications for public order… Athens had taken the first steps on the road to being a state governed not only by rules, but by fair rules” (2015, p. 151).

Solon’s reforms were strengthened and reconfigured by Cleisthenes, who rose to power as a result of a mass uprising. Cleisthenes continued Solon’s political reforms and further strengthened judicial institutions in Athens. He also expanded state capacity and introduced various public services, financed for the first time by an elaborate fiscal system. Many of his reforms increased equality before the law. Particularly important in this respect was the introduction of ostracism. Every year at a pre-specified date the Athenian Assembly could take a vote on whether to ostracize someone. If at least 6,000 people voted in favor of an ostracism, then each Athenian citizen could write the name of whomever they wanted to ostracize on a shard of pottery (called an ostrakon, hence the term ostracism). The person whose name was written on the most shards was ostracized and exiled from Athens for 10 years, under penalty of death. Athenian ostracism thus involved a subtle mix of legal and community enforcement. In practice, it was directed only towards elite Athenians: the roughly 15 Athenians known to have been ostracized in the Classical period were all famous statesmen, including Themistocles, the mastermind of the Athenian victory over the Persians at Salamis and one of the most powerful Athenians at the time. Ober concludes his discussion of ostracism by writing:

“Building on the laws of Solon, the Cleisthenic package changed the rules of the Athenian state in ways that seems to have oriented individual behavior of elite and ordinary citizens alike in an overall growth-positive direction” (2015, p. 175).
What explains these moves towards, by the standards of the time, unparalleled equality before the law? We argue that our theory provides one aspect of the answer: this was a period during which the ability of the elites to use coercion against regular Athenians declined (Snodgrass, 1990, Ober, 2015). An important reason for this social change was the interplay between developments in military technology and the nature of Athenian society. During the time of the “palace economies” of bronze age Greece, around 16-11th centuries BC, most weapons were made of bronze. Bronze was very expensive, and consequently weapons were monopolized by the elite. Military technology changed in the intervening centuries, especially with the introduction of iron weapons, which started being used by Athenian citizen infantry—“hoplites”. In the famous words of Gordon Childe (1942), “Iron democratized agriculture and industry and warfare too.” Ober explains this as follows:

“In the Iron Age, given the right social will on the part of the community’s leaders, it was relatively easy for many local men to be outfitted with the basic infantry equipment of iron-tipped spear, wooden shield, and headgear—the equipment that was eventually elaborated and canonized in the ‘hoplite panoply’. The ready availability of iron thus made it more difficult for Iron Age elites to monopolize the potential for organized violence.” (2015, p. 130).

This democratization of warfare tilted the balance of power away from the elites towards the citizens. In the context of our model, it is an example of technological change limiting the extent of coercive punishments. Our key comparative static then implies that the resulting social changes should lead to greater equality before the law.

This interpretation is bolstered by a comparison of Athens to its rival state of Sparta. Sparta was affected by the same technological changes and underwent some major social and political changes. In Sparta too a system of relative equality among citizens developed. Spartan citizens, for example, are referred to as homoioi (equals) and had various rights (even if in reality there was quite a bit of inequality among them). But the differences between Athens and Sparta were stark. The homoioi were the minority and made up the warrior elite class, where all males over the age of 20 were organized in messes and subjected to continuous and rigorous military training. The rest of society comprised a large mass of helots, who were slaves, or at the very least state-owned serfs, and another type of servile labor, the perioikoi, who were settled in villages and specialized in manufacturing and especially weapon making for Spartan citizens. Notably, helots and perioikoi had very limited rights in the highly hierarchical Spartan society. Violence against helots was commonplace and sometimes sanctioned by the Spartan state. This contrasts sharply with the protections of slaves’ rights in Athenian society mentioned above. The number of slaves in Athens was also far less than the number of helots in Sparta. Undergirding this class-based structure and social hierarchy was a high degree of economic inequality (see, e.g. Ober, Chapters 5 and 6).

This structure of society, together with the fact that weapons were directly procured by the
state, was a major difference between Athens and Sparta. Importantly, the democratizing role of iron weapons emphasized by Childe played a more limited role in Sparta, and only to the extent that it empowered the small minority of Spartan warrior-citizens. This may be explained both by the state’s direct control over the means of coercion and by the observation that the very large number of helots created a very different type of social conflict in Sparta.

7.2 Equality before the Law in Britain

The British case is often emphasized in discussions of the rule of law, with many scholars tracing the roots of these notions to the Middle Ages or even earlier. These important legal and political traditions notwithstanding, Britain remained far from equality before the law as late as the mid-19th century. An emblematic example is provided by a set of laws creating onerous obligations for manual workers and privileges for employers, who could ban workers from quitting their jobs, or even from turning down unattractive offers (Steinfeld, 2001, Naidu and Yuchtman, 2013).

The Statute of Laborers, enacted in the 14th century, empowered landowners to compel workers to work at set wages. In Steinfeld’s words, “The English laboring poor of this period…were subject to an oppressive regime of legal regulation” (p. 8), and “In the 14th and 15th centuries, justices regularly ordered imprisonment for those who violated their oral employment agreements by departing before the term agreed” (p. 28). This law was reconfirmed by later, 16th-century statutes, and was extended to a handful of artisanal occupations in the 18th century. It was also imported by the American colonies and formed the core of their labor law.

As the demand and competition for labor in industry increased in the course of the Industrial Revolution, there were calls for these laws to be extended to industrial workers. The 1823 Master and Servant Act applied similar provisions to all manual workers, enabling employers to prosecute their workers for contract breach if they quit their jobs or did not accept the proffered contract terms. Prosecutions under the act were very common, and while fines were the standard penalty, whippings and imprisonments were also frequent.

However, by this time powerful social and political changes were already making the disproportionate power of employers over laborers less tenable. This process had started in the 17th century. Steinfeld notes, “It was during the 17th century that the English tradition of invoking ‘ancient native liberties’ and ‘rights of the freeborn’ first became an important feature of the Anglo-American political landscape” (p. 94), and these arguments began to be used to assert the rights of workers. The Levellers argued in 1646, for example, that “As God created every man free in Adam, so by nature all are alike free men born” (p. 95). By the 18th century, the prevailing values in English society had undergone a transformation, and “As common people agitated to secure the blessings of liberty for themselves, the old assumption that labor agreements convey property came under greater and greater scrutiny” (p. 106). These currents and the resistance to coercive labor
institutions gathered steam in the 19th century as political participation started broadening with
the First Reform Act of 1832 and workers became better organized into trade unions, and they
ultimately triggered a process of gradual judicial reform.

A first response to these social pressures was the 1867 Master and Servant Act, which prohibited
whippings and imprisonment, even as it simultaneously increased the ability of magistrate courts to
compel workers to work at the terms offered by their employers. While it superficially strengthened
aspects of Britain’s unequal labor regime, the 1867 act in fact marked the beginning of this regime’s
end. Crucially, once coercive punishments were taken off the table, the ability of employers to use
their privileges before the law was quickly curtailed. The act itself was finally repealed in 1875 in
a critical step towards equality before the law and a broader set of rights for British laborers.

7.3 Defensive Modernization and Equality before the Law in Japan

Several cases of defensive modernization triggered by external threats illustrate our second key
comparative static result—the response of equality before the law to an increase in the return to
effort. Let us start with 19th-century Japan.

Japan in the first half of the 19th century was governed by the Tokugawa Shogunate, which set
up a highly hierarchical, class-based, quasi-feudal society ruled by the Tokugawa Shogun. At the
top of the hierarchy was the Shogun, residing in Edo (now Tokyo), the emperor (who still existed
as a figurehead), his courtiers, and the daimyo, who were local lords governing the 300 regions
of Japan at this time. Below this top layer were the samurai who made up the warrior class in
Japan and had various special social and legal privileges, most importantly the exclusive right to
bear arms. At the bottom of the hierarchy were peasants, craftsmen and merchants, who had a
lower legal standing and were not allowed to bear arms. This hierarchy was supported both by
the imbalance of coercive power between non-elites (often viewed with suspicion by the Tokugawa
state) and the samurai elite and their masters, and by customs and social norms.

Though this rigid system created obvious advantages for the samurai and the landowning elite,
it also kept Japan technologically and economically backward, a problem that was laid bare when
Commodore Matthew C. Perry sailed into the Bay of Tokyo in 1853–54 and forced Japan to open
to foreign (especially American) trade. Even before the arrival of Commodore Perry, tensions
in Tokugawa Japan were obvious and there were elite-driven moves for reform, especially after
Japanese elites witnessed the Chinese empire crumble in the First Opium War. These movements
gained urgency and power with the threat posed by Perry’s fleet. They ultimately convinced
various factions of the Japanese elites to undertake an ambitious reform program as a defensive
modernization strategy to strengthen Japanese state institutions against foreign threats.

The key event was the Meiji Restoration of 1866, which disbanded the Tokugawa system and
“restored” the powers of the emperor. The Restoration initiated a process of military, social and
economic modernization (Jansen, 2002, Ravina, 2017). The Meiji government removed the de jure unequal treatment of different social classes and disarmed the samurai, some of whom remained specialists in coercion, but now as police officers under the control of the central state. It also created a professional army and navy. In Ravina’s words:

“The creation of the Meiji army and navy was an explicit rejection of Tokugawa social and political traditions. Since the late 1500s, Japanese rulers had separated warriors from commoners; commoners were effectively disarmed, while samurai were distinguished by their rights to carry two swords. . . . Hereditary warriors were no longer needed. Instead, Japan’s new military would comprise Japanese subjects from all classes” (2017, p. 5).

A watershed was the Meiji Constitution, drafted in the 1880s and finally promulgated in 1890. The constitution introduced notions such as due process, freedom of movement, freedom of speech, and private property for all Japanese. While 19th-century Japan remained an oligarchic society, these changes created a much greater degree of equality before the law.

An important question in this context is why the Meiji reforms did not simply modernize the military and the fiscal system, but also took fairly major steps towards legal equality. Our second key comparative static suggests a potential answer. The trigger for these reforms—mounting external threats—corresponds in our framework to a rise in the marginal returns to effort. Our analysis predicts that such shocks lead to greater equality before the law, because they complement increased effort by citizens and elites, which became more critical for the survival of the country. This is consistent with Ravina’s emphasis (2017, p. 152) that these reforms were essential for “creating a single, unified Japanese nation”, capable of holding its own in international affairs.

7.4 Other Examples of Defensive Modernization

There are many parallels between the Japanese experience and other well-known examples of defensive modernization, which can also be interpreted as instances of reforms triggered by external threats that raised the returns to societal effort and cooperation. As in the Japanese case, these reforms not only targeted the military and the fiscal system, but also introduced some degree of equality before the law.

The process of reform in 19th-century Prussia, which is often interpreted as a response to the threat posed by the mass armies organized by the French Revolution and Napoleon, was also a major step towards equality before the law. Before the 19th century, Prussia was a highly hierarchical society where large fraction of the population (especially in the East) did not enjoy legal protection or many rights. The process of reform disbanded various vestiges of serfdom and recognized all Prussians as citizens (Fisher, 1903, Blanning, 1989, Acemoglu et al., 2011). This process was a
first step towards the unification of Germany as well, and (as in the Japanese case) was intended
to create a unified German nation, in part as a defensive modernization strategy.

Another example is provided by the Tanzimat reforms in the Ottoman Empire, promulgated in
the Rose Garden Edict in 1839. This too was a process of top-down reform motivated by the relative
decline of the Ottoman state and army compared to its European rivals, and by a number of high-
profile military defeats suffered by Ottoman forces. Once again, the reforms did not just modernize
the army, but initiated sweeping societal changes. Most importantly from our perspective, for the
first time in Ottoman history some degree of equality before the law was introduced, including for
various non-Muslim minorities (Zürcher, 2004, Owen, 2004).

8 Private Benefits of Cooperation

We have assumed thus far that effort is a pure public good—it creates equal benefits for everyone
in society. Though this assumption is a natural starting point and substantially simplifies our
analysis, it is also useful to go beyond it for at least two reasons. First, while many forms of pro-
social behavior generate benefits for everybody in society, these benefits are not necessarily equally
distributed. For example, effort directed at production may benefit everyone who consumes the
relevant good or uses it as an input (especially when markets are not perfectly competitive), but
may generate even greater benefits for one’s business partners or associates. Second, the pure
public good nature of cooperation implies that elites can be favored only by having to exert less
effort than normal agents. In practice, elites may also receive special treatment from the non-elites
who interact with them more closely (as their employees, servants, serfs, etc.). In this section, we
generalize our baseline environment to address these issues.

Specifically, we analyze the random matching model described at the end of Section 2, where
effort decisions are taken non-anonymously. In every period, agents first randomly match in pairs
and observe their partner’s status (normal or elite) and then exert effort (which disproportionately
benefits one’s partner), and then each elite agent has the option of punishing her partner. Note
that any equilibrium of this non-anonymous random matching model in which players do not
condition their effort choices on their partners’ status reduces to an equilibrium of the anonymous
random matching model—and thus an equilibrium of our baseline, centralized model—so the non-
anonymous random matching model is effectively a generalization of the baseline model. In this
section, we show how this generalization affects the structure of incentives, and we establish that our
most important comparative static result generalizes to this environment: a reduction in coercive
capacity $g$ encourages greater equality before the law.$^{30}$

---

$^{30}$Our other comparative static results do not generalize without further conditions. These results are all robust
to introducing a small private goods component to cooperation, but when the private goods component is large the
results become more nuanced. The issue is that each type of agent chooses different effort levels when matched with
normal and elite agents, and it is difficult to rule out these two effort levels moving in opposite directions with respect
To model the fact that cooperation imposes positive externalities on society without being a pure public good, we assume that a fraction \(1 - \lambda \in [0, 1]\) of the benefits of cooperation accrue only to one’s partner rather than to society at large. Thus, \(\lambda = 0\) corresponds to pure private goods (i.e., cooperation generates no positive externalities), and \(\lambda = 1\) corresponds to pure public goods (and is thus identical to our baseline environment). Formally, when player \(i\) chooses effort \(x_i\), her partner chooses effort \(x_j\), and the distributions of effort levels among normal agents and the elite are, respectively, \(F_N\) and \(F_E\), player \(i\)’s stage payoff is

\[
(1 - \lambda) f_N (x_j) + \lambda ((1 - \alpha) \mathbb{E}_{F_N} [f_N (x)] + \alpha \mathbb{E}_{F_E} [f_E (x)]) - x_i
\]

if her partner is normal, and

\[
(1 - \lambda) f_E (x_j) + \lambda ((1 - \alpha) \mathbb{E}_{F_N} [f_N (x)] + \alpha \mathbb{E}_{F_E} [f_E (x)]) - x_i
\]

if her partner is elite. A (symmetric, stationary, subgame perfect) equilibrium is now parameterized by four variables, \((w, x, y, z)\), where \(w\) is a normal agent’s equilibrium effort when matched with another normal agent, \(x\) is a normal agent’s effort when matched with an elite, \(y\) is an elite’s effort when matched with a normal agent, and \(z\) is an elite’s effort when matched with another elite.

Our main result in the private goods model is that increasing coercive capacity decreases equality before the law, which again implies that limits on the extent of coercion are one major factor leading to the emergence of equality before the law.

**Proposition 12** Under endogenous equality before the law, suppose the elite-optimal level of equality before the law \(\rho^*\) is strictly less than 1. Then the solution to the elites’ problem is differentiable in \(g\), and \(dw^*/dg \geq 0\), \(dx^*/dg \geq 0\), \(dy^*/dg \leq 0\), \(dz^*/dg \leq 0\), and \(d\rho^*/dg \leq 0\).

The basic intuition for this result is similar to that in our baseline model, in particular Proposition 6, though the proof (deferred to the Online Appendix) is more complicated as there are now four on-path effort levels, rather than two as in the baseline model. Nevertheless, as in our baseline environment, an increase in \(g\) relaxes normal agents’ incentive compatibility constraints and allows elites to demand greater effort from normal agents both when normal agents match with each other and when they match with elites. As there are diminishing returns to effort in each match, this reduces elites’ returns from raising their own effort in order to encourage yet greater effort from normal agents. Hence, elites work less in the elite-optimal equilibrium when \(g\) is higher, and therefore have less need to subject themselves to the law.

---

33
9 Heterogeneous Elites

Finally, we consider two extensions of our framework that allow for heterogeneity—in terms of productivity and political power—within the elite. We investigate what types of changes in the composition of the elite encourage greater equality before the law.

9.1 Heterogeneous Productivity within the Elite

Several historical cases of the expansion of equality before the law have been attributed to shifts in political power among subsets of the elite with heterogeneous economic interests. Most notably, it is often argued that several aspects of economic and social modernization in late-medieval Western Europe resulted from the changing political balance between different segments of the elite, in particular between commercial and landed interests (Moore, 1966, Aston and Philpin, 1987). We now show that in a simple extensions of our model, a shift of political power away from landed interests (here interpreted as the less productive part of the elite) to more productive commercial interests can support the emergence of equality before the law.

We assume there are two elite types that differ according to a productivity parameter $b$: an elite agent with productivity $b$ who exerts effort $y$ generates output $f_E(by)$. Fraction $\alpha_H$ of the population are (high-productivity, commercial) elites with productivity $b_H$, and fraction $\alpha_L$ of the population are (low-productivity, landed) elites with productivity $b_L \leq b_H$. We assume that an individual’s elite status and output are observable, but her productivity is unobservable. Thus, members of the two elite subgroups cannot be asked to produce different output levels, since otherwise each would pretend to be a member of the group that produces less output.\footnote{This is one part of our analysis that does depend on the continuum population assumption: the claim in the text is clearly true with a continuum, but would require more careful justification with a finite population.}

If the equilibrium effort level of high-productivity elites is $y_H$, then the equilibrium effort level of low-productivity elites is $(b_H/b_L)y_H$. Noting that all elites produce output $f_E(b_Hy_H)$, the resulting incentive constraints are

\begin{align}
x & \leq \delta [(1 - \alpha_H - \alpha_L) f_N(x) + (\alpha_H + \alpha_L) f_E(b_Hy_H)] + g \\
y_H & \leq \delta [(1 - \alpha_H - \alpha_L) f_N(x) + (\alpha_H + \alpha_L) f_E(b_Hy_H)] + pg \\
b_H/y_H & \leq \delta [(1 - \alpha_H - \alpha_L) f_N(x) + (\alpha_H + \alpha_L) f_E(b_Hy_H)] + pg.
\end{align}

As $b_H > b_L$, the second constraint is slack and can be dropped. We are thus back to a problem with two constraints, and now the elites’ incentive constraint can be assumed to bind and is used to define the elite-optimal level of equality before the law.

Our main goal in this subsection is to investigate the implications of a shift in political power from less productive to more productive elites. To model this in the simplest possible way, we...
assume that negotiations within the elite lead to the maximization of a weighted average utility of the two elite groups, with (Pareto) weight $\beta$ on high-productivity elites and $1 - \beta$ on low-productivity elites. The effort level of the elite is then determined as a solution to the following maximization problem:

$$\max_{y_H \geq 0} \left(1 - \alpha_H - \alpha_L\right) f_N \left(x^* \left(y_H\right)\right) + \left(\alpha_H + \alpha_L\right) f_E \left(b_H y_H\right) - \left(\beta + (1 - \beta) \frac{b_H}{b_L}\right) y_H,$$

where $x^* \left(y_H\right)$ is implicitly defined as the level of $x$ that binds the normal agents’ incentive constraint. Implicitly differentiating $x^* \left(y_H\right)$, we obtain

$$\frac{dx^*}{dy_H} = \frac{\delta \left(\alpha_H + \alpha_L\right) b_H f'_E \left(b_H y_H\right)}{1 - \delta \left(1 - \alpha_H - \alpha_L\right) f'_N \left(x\right)}.$$

Using this expression, the first-order condition with respect to $y_H$ can be written as

$$\frac{\left(\alpha_H + \alpha_L\right) b_H f'_E \left(b_H y_H\right)}{1 - \delta \left(1 - \alpha_H - \alpha_L\right) f'_N \left(x\right)} = \frac{b_H}{b_L} - \beta \frac{b_H - b_L}{b_L}.$$

The right-hand side is decreasing in $\beta$. Moreover, as $f_N$ and $f_E$ are concave and $x$ is increasing in $y_H$, the left-hand side is decreasing in $y_H$. Hence, $x^*$ and $y^*_H$ are increasing in $\beta$. Finally, $\rho^*$ is defined to satisfy (13) with equality, and therefore

$$\frac{d\rho^*}{d\beta} \left(\begin{array}{l}
\frac{b_H}{b_L} \frac{dy_H}{d\beta} - \delta \left[1 - \alpha_H - \alpha_L\right] f_N \left(x^* \left(y_H\right)\right) + \left(\alpha_H + \alpha_L\right) f_E \left(b_H y_H\right) \right] \\
= \frac{b_H}{b_L} \frac{dy_H}{d\beta} - \delta \left[\beta + (1 - \beta) \frac{b_H}{b_L}\right] \frac{dy_H}{d\beta} \geq 0,
\end{array}\right.$$

where the second equality follows by the first-order condition with respect to $y_H$.

In sum, an increase in the political power of the more productive elite group (loosely approximating commercial interests in late middle-age Europe) is likely to lead to an increase in equality before the law. The intuition is that an increase in equality before the law raises the level of output required of all elite agents, and generating this increased output is less costly for more productive elites. Since the marginal benefit of an increase in equality before the law is the same for all elites while the marginal cost of an increase in equality before the law is less for more productive elites, an increase in effort level is relatively more beneficial for more productive elites. Hence, the more politically powerful are the more productive elites, the greater is the equilibrium elite effort, and this translates into a greater level of equality before the law.
9.2 Enforcement Hierarchy

Suppose again that there are two elite groups, now corresponding to minor elites (say barons) and more powerful, major elites (say dukes). These two groups are now equally productive but differ in their vulnerability to coercion. Specifically, suppose that—in the absence of equality before the law—normal agents are vulnerable to coercion from both types of elites, while minor elites (type 1 elites) can be coerced by more powerful elites (type 2 elites), and the latter are initially completely immune to coercion. The level of equality before the law now parameterizes both the vulnerability of minor elites to coercion from other minor elites and the vulnerability of major elites to coercion from both minor and major elites. The resulting incentive constraints are

\[
\begin{align*}
x & \leq \delta \left[ (1 - \alpha_1 - \alpha_2) f_N(x) + \alpha_1 f_E(y_1) + \alpha_2 f_E(y_2) \right] + (\alpha_1 + \alpha_2) g \\
y_1 & \leq \delta \left[ (1 - \alpha_1 - \alpha_2) f_N(x) + \alpha_1 f_E(y_1) + \alpha_2 f_E(y_2) \right] + (\rho \alpha_1 + \alpha_2) g \\
y_2 & \leq \delta \left[ (1 - \alpha_1 - \alpha_2) f_N(x) + \alpha_1 f_E(y_1) + \alpha_2 f_E(y_2) \right] + \rho (\alpha_1 + \alpha_2) g.
\end{align*}
\]

Intuitively, as \( \rho \) increases, this closes both the gap in privilege between normal agents and elites as a whole and the gap between the minor and major elites.

With two different elite incentive constraints, the issue of what level of equality before the law is optimal for the elites is delicate. For example, if either minor elites or major elites could choose both \( \rho \) and the resulting equilibrium, they would choose \( \rho = 1 \) while requiring more effort from the other elite group. These unintuitive possibilities disappear when all three incentive constraints bind, and this is the case on which we focus in this subsection.\(^\text{32}\) This focus thus rules out equilibria where one elite group sets a high level of equality before the law to coerce the other elite group while exerting low effort itself. Consequently, for any value of \( \rho \), the two elite groups differ in their vulnerability to coercion, but the full force of the level of \( \rho \) that is chosen applies to both groups.

Under the assumption that all incentive constraints bind, consider again the problem of a planner with Pareto weights \((\beta, 1 - \beta)\) on the two elite groups. When all incentive constraints bind, this problem involves only the single choice variable \( \rho \). Letting \( x(\rho) \), \( y_1(\rho) \), and \( y_2(\rho) \) be the resulting effort levels, we can obtain the endogenous level of equality before the law as a solution to the following problem:

\[
\max_{\rho \in [0,1]} (1 - \delta) \left[ (1 - \alpha_1 - \alpha_2) f_N(x(\rho)) + \alpha_1 f_E(y_1(\rho)) + \alpha_2 f_E(y_2(\rho)) \right] - \beta (\rho \alpha_1 + \alpha_2) g - (1 - \beta) \rho (\alpha_1 + \alpha_2) g.
\]

It is straightforward to see that this objective function is supermodular in \((\beta, \rho)\). Hence, the set

\[^{32}\text{The assumptions that } \delta f_N'(0) > 1 \text{ and } \delta f_E'(0) > 1 \text{ and that these functions are concave and bounded imply that there exists a positive vector } (x_1, y_1, y_2) \text{ where all three constraints bind. The concavity of these functions also implies that there is only one such vector.}\]
of optimal values of $\rho$ is increasing in $\beta$ in the strong set order. Thus, when minor elites have more political power, the resulting level of equality before the law is higher. The intuition is that since minor elites are already exposed to coercion by major elites, greater equality before the law increases the effort of major elites by relatively more than it increases the effort of minor elites. This makes minor elites more inclined to favor equality before the law. Thus, an increase in minor elites’ political power leads to greater equality before the law.

This comparative static result, like the one with respect to $\alpha$ discussed above, is related to North, Wallis and Weingast’s (2009) argument that rule of law among the elite is a precursor to the emergence of equality before the law for all agents. Consistent with this comparative static (and with North, Wallis and Weingast), several historical episodes support the notion that political changes that strengthen minor elites encourage greater equality before the law. For example, the Magna Carta was an agreement imposed by barons on King John in 1215, limiting his powers and ability to act without the approval of the barons. But the final charter was formulated as a concession from the king “to all the free men of our kingdom”, and went so far as to restrict the ability of landowners to impose forced labor on their own serfs (see Holt, 2015, and the discussion in Acemoglu and Robinson, 2019). Our extension in this subsection is a simple formalization of these ideas: as the political power of minor elites increases relative to that of more powerful elites, this encourages an extension of equality before the law for all agents in society.

10 Conclusion

This paper is a first step towards developing a theory of the rule of law. It focuses on the emergence of a vital component of rule of law—equality before the law. Our approach is to model the organization of society via a repeated game in which cooperation and public good provision need to be encouraged. One way of doing this—reminiscent of the organization of stateless societies—is by community enforcement, relying only on the carrot of future cooperation: agents that exert the requisite amount of effort benefit from future cooperation, and those that deviate are excluded from these benefits. Another way of organizing society is to combine this carrot with the stick of coercion, which directly imposes costly punishments on those who deviate from laws or social norms. We assume that, as has almost always been the case in history, centralized states are initially under the control of a subset of privileged agents—the “elite”—and coercive punishments favor this group. In contrast to the low levels of coercion and limited inequality that prevail under community enforcement, under elite enforcement coercion and inequality are high, supported by laws and norms that both benefit the elite. Elites are “above the law” in the precise sense of not being subject to coercion. Potentially shedding light on some major debates in anthropology, we show that the transition from community enforcement to elite enforcement can increase or decrease the welfare of normal agents: on the one hand, it encourages greater productive effort; on the other,
it privileges elites at the expense of normal agents.

The most important part of our analysis concerns situations where the elite can choose between elite enforcement and various degrees of equality before the law. We show that it may be optimal—even from the viewpoint of the elite—to introduce full equality before the law, which combines high coercion with low inequality. Norms and laws are again interdependent: equal treatment before the law coincides with a change in norms towards expectations of identical behavior from elites and from normal agents. The key mechanism is that by stripping the elite of their privileges, equality before the law enhances the carrot of future cooperation for normal agents. This encourages normal agents to exert greater effort, which can benefit everyone in society, including elites.

We identify several factors that encourage the emergence of equality before the law. Perhaps the most important is a decline in the extent of coercive punishments that elites can impose on citizens. Such a change in the technology of coercion can arise for several reasons, ranging from equalizing changes in military technology, to increased political power of citizens, to social changes that make certain harsh punishments simply unacceptable. The intuition for this central comparative static is that when punishments are limited, the stick of coercion becomes less attractive compared to the carrot of cooperation, which tilts society towards greater effort from the elite, and thus towards greater legal equality. A direct increase in the political power of normal agents has similar effects. We also establish that an increase in marginal returns to effort (but not average returns) favors equality before the law. This can be interpreted as a national emergency or a change in international circumstances necessitating greater cooperation and investment in public goods, such as the defensive modernization in 19th-century Prussia, Japan, or the Ottoman Empire. Exploring the role of economic inequality, we show that when elites become richer, this may discourage them from exerting additional effort and thus hinder legal equality. When elites are heterogeneous in terms of their economic investments and productivity (e.g., divided between landowners and commercial interests), a strengthening of more productive segments of the elite also leads to greater equality before the law. Finally, consistent with the emphasis of North, Wallis and Weingast (2009), we show that various changes encouraging “rule of law among the elite”—resulting either from an increase in the size of the elite or a change in the balance of power within a heterogeneous elite towards its weaker members—likewise favor greater equality before the law.

Many interesting areas remain to be explored. Possible extensions of our model include endogenizing the size of the elite (for example, by introducing some amount of social mobility, which could itself be determined as part of the equilibrium) and allowing the elite to choose their coercive capacity. It could be fruitful to apply similar ideas to the internal organization of firms. A key aspect of organizations that has received relatively little attention from economists is the balance of power between management and workers. Tilting this balance in a way that induces managers to exert more effort can incentivize workers, either via repeated game incentives or gift-exchange type considerations. The analogue of our comparative static with respect to coercive capacity here
might be studied by considering changes in societal values, social norms, and institutions that make it less acceptable for managers to ask for certain actions from their employees. Interesting issues in this context might include the effect of exit options and markets on internal organization, as well as what aspects of firm architecture affect the balance between management and workers. Yet another direction in this context might be to merge a model of labor coercion as in Acemoglu and Wolitzky (2011) with repeated game considerations, so that the carrot of future cooperation interacts with coercive behavior by employers.

The traditional notion of rule of law emphasizes not only equality before the law, but also effective legal constraints on executive power—the sovereign must be bound by the law. Modeling this aspect of the rule of law together with equality before the law is an important area for future theoretical research. Yet another critical role of the law is conflict resolution; a particularly interesting issue here is the emergence of legal equality in conflict resolution. Finally, Hayek (1960) emphasizes the importance of the gradual evolution over time of the rule of law, an idea which is echoed by many legal philosophers, including Hart (1961). A challenging but important area for future research is to systematically investigate this issue (i.e., whether there are reasons for gradual, evolutionary changes to support the rule of law, and more generally reasons for laws to be consistent with existing norms and customs). Relatedly, our approach has abstracted from the fact that, to be effective, laws need to be obeyed, which may also require them to be consistent with norms (e.g., Acemoglu and Jackson, 2017) or to have legitimacy in the eyes of the public (e.g., Tyler, 2006). It would be fruitful to investigate how these issues interact with equality before the law. Last but not least, empirical research directed at understanding the causes and implications of the emergence of equality before the law is another important area for subsequent research.

Appendix: Proofs

Proof of Proposition 1

If \( x \) is an equilibrium effort level, then

\[
f_N(x) - x \geq (1 - \delta) f_N(x).\]

This follows as the left-hand side is the equilibrium payoff, and the right-hand side is a player’s payoff from deviating to \( x_i = 0 \) and subsequently receiving her minmax payoff (under community enforcement) of 0. Hence, \( x \leq \delta f_N(x) \) in every equilibrium, and therefore (as \( f_N \) is concave) \( x \leq \bar{x}^{CE} \). Conversely, grim trigger strategies can support any effort level up to \( \bar{x}^{CE} \) as an equilibrium.

Proof of Proposition 2

If \((x, y)\) are equilibrium effort levels, then

\[
(1 - \alpha) f_N(x) + \alpha f_E(y) - x \geq (1 - \delta) [(1 - \alpha) f_N(x) + \alpha f_E(y)] - g.
\]
This follows as the left-hand side is a normal agent’s equilibrium payoff, and the right-hand side is a normal agent’s payoff from deviating to \( x_i = 0 \) and then being minmaxed, noting that a normal agent’s minmax payoff is \(-g\) because of coercive punishments. Rearranging this expression yields (3). The argument for (4) is the same, except that an elite agent’s minmax payoff is 0 rather than \(-g\). Moreover, (3) and (4) are sufficient as well as necessary for \((x, y)\) to be a pair of equilibrium effort levels, because under these conditions grim trigger strategies combined with coercive punishment of any deviator in every period following the deviation support constant effort at \( x \) and \( y \) for normal and elite agents, respectively. Finally, it is clear that (3) binds at the optimum, as increasing \( x \) increases the objective and also relaxes constraint (4).

For the last part of the result, let \( x^*(y) \) be the value of \( x \) that binds (3). By the implicit function theorem,

\[
\frac{dx^*(y)}{dy} = \frac{\delta \alpha f'_E(y)}{1 - \delta (1 - \alpha) f'_N(x^*(y))}.
\]

The total derivative of the objective with respect to \( y \) is then equal to

\[
(1 - \alpha) f'_N(x^*(y)) \frac{dx^*(y)}{dy} + \alpha f'_E(y) - 1 = \frac{\alpha f'_E(y)}{1 - \delta (1 - \alpha) f'_N(x^*(y))} - 1.
\]

By complementary slackness, at the solution either (i) \( y = 0 \) and the derivative is non-positive; (ii) \( y > 0 \), (4) is slack, and the derivative equals 0; or (iii) constraint (4) binds and the derivative is non-negative. This argument yields (5)–(7).

**Proof of Proposition 3**

As \( f_N \) is concave and \( x^{EE} = \delta \left[(1 - \alpha) f_N(x^{EE}) + \alpha f_E(y^{EE})\right] + g \), we have that \( \delta (1 - \alpha) f'_N(x^{EE}) < 1 \) uniformly over \( \alpha \). By (5)–(7) and \( f'_E(0) < \infty \), there exists \( \hat{\alpha} > 0 \) such that if \( \alpha < \hat{\alpha} \), then \( y^{EE} = 0 \) for all \( g > 0 \). Hence, for \( \alpha < \hat{\alpha} \), \( dx^{EE}/dg \geq 0 \) (as \( x^{EE} \) is defined as the solution to \( x = \delta (1 - \alpha) f_N(x) + g \)), and

\[
\frac{du^{EE}_N}{dg} = \left((1 - \alpha) f'_N(x^{EE}) - 1\right) \frac{dx^{EE}}{dg}.
\]

So there exists \( \hat{x} \) such that \( du^{EE}_N/dg \) is non-negative for \( x^{EE} < \hat{x} \) and non-positive for \( x^{EE} > \hat{x} \). Again using the fact that \( dx^{EE}/dg \geq 0 \), we see that \( u^{EE}_N \) is single-peaked in \( g \). Since \( u^{CE}_N \) does not depend on \( g \), the result follows.

**Proof of Proposition 4**

If the solution to the elites’ problem involves \( \rho^* = 0 \), the problem reduces to that under elite enforcement. If instead \( \rho^* > 0 \), then (10) binds by the assumption that \( \rho^* \) is minimal. As (9) always binds, when \( \rho^* = 1 \) it immediately follows that \((x^{EL}, y^{EL}) = (\hat{x}^{EL}, \hat{y}^{EL})\). When \( \rho^* \in (0, 1) \), the elite-optimal equilibrium is an interior solution to (8), subject to \( y < \hat{y}^{EL} \). Hence, \( y^{EL} \) must satisfy the first-order condition (11) derived in the proof of Proposition 2.

---

\[^{33}\text{The denominator is non-zero because, by concavity of } f_N \text{ and inspection of (3), } 1 - \delta (1 - \alpha) f'_N(x) \text{ must be strictly positive at } x = x^*(y).\]
Proof of Proposition 5

First, note that \((x^{EL}, y^{EL}) \geq (x^{EE}, y^{EE})\), with strict equality if \(\rho^* > 0\). To see this, note that \(x^{EE}\) is the positive root of the concave function

\[
\delta [(1 - \alpha) f_N (x) + \alpha f_E (x - g)] + g - x,
\]

and when \(\rho^* > 0\), \(x^{EL}\) is the positive root of the concave function

\[
\delta [(1 - \alpha) f_N (x) + \alpha f_E (x - (1 - \rho^*) g)] + g - x
\]

(where we have used the fact that \(y^{EL} = x^{EL} - (1 - \rho^*) g\) when \(\rho^* > 0\)). The latter function is everywhere strictly greater than the former, so its positive root is strictly greater. The argument for \(y^{EL} \geq y^{EE}\) is similar.

Next, as normal agents’ incentive constraint binds, we have

\[
u^{EE}_N = (1 - \delta) [(1 - \alpha) f_N (x^{EE}) + \alpha f_E (y^{EE})] - g,
\]

\[
u^{EL}_N = (1 - \delta) [(1 - \alpha) f_N (x^{EL}) + \alpha f_E (y^{EL})] - g.
\]

As \(x^{EL} \geq x^{EE}, y^{EL} \geq y^{EE}\), and \(f_N\) and \(f_E\) are increasing, it follows that \(u^{EL}_N \geq u^{EE}_N\).

Finally, we have seen that if \(\rho^* = 1\) then \(x^{EL} = y^{EL}\), and hence \(u^{EL}_N = u^{EE}_N\). Since \(u^{EL}_E \geq u^{EE}_E \geq u^{CE}\), with strict equality if \(\rho^* > 0\), it follows that \(u^{EL}_N > u^{CE}\). ■

Proof of Proposition 6

Let \(u_E (g)\) denote the value of (12) given coercive capacity \(g \geq 0\). We claim that \(u_E (g)\) is a strictly increasing and strictly concave function of \(g\). Strict monotonicity is obvious, as one possible response to an increase in \(g\) is to increase \(x\) while leaving \(y\) unchanged. For strict concavity, suppose \((x, y)\) is a solution given coercive capacity \(g\) and \((x', y')\) is a solution given coercive capacity \(g' > g\). By strict monotonicity, \((x, y) \neq (x', y')\). Moreover, for all \(\beta \in (0, 1)\), \((x^*, y^*) = (\beta x + (1 - \beta) x', \beta y + (1 - \beta) y')\) is feasible given coercive capacity \(\beta g + (1 - \beta) g'\) (as \(f_N\) and \(f_E\) are concave), and elite utility at \((x^*, y^*)\) is strictly greater than the \(\beta\)-weighted average of elite utility at \((x, y)\) and \((x', y')\).

Next, let \(\mu_N\) be the Lagrange multiplier on (3). Note that

\[
du_E (g) = \mu_N.
\]

Hence, \(\mu_N\) is strictly decreasing in \(g\).

It is now straightforward to show that (elite-optimal) normal agent effort is nondecreasing in \(g\) and elite agent effort is nonincreasing in \(g\). In particular, the first-order conditions of the Lagrangian with respect to \(x\) and \(y\) are

\[
(1 - \alpha) f'_N (x) = \mu_N (1 - \delta (1 - \alpha) f'_N (x)),
\]

\[
1 - \alpha f'_E (y) = \mu_N (1 - \delta x f'_E (y)).
\]

At an interior optimum, \(\delta (1 - \alpha) f'_N (x) < 1\) and \(\alpha f'_E (y) < 1\). As \(f_N\) and \(f_E\) are concave, implicitly differentiating the first-order conditions and using the fact that \(\mu_N\) is strictly decreasing implies that the optimal value of \(x\) is strictly increasing, and the optimal value of \(y\) is nonincreasing and is strictly decreasing when \(\delta > 0\).
Finally, to derive the comparative static on \( \rho^* \), recall that \( \rho^* \) is the value of \( \rho \) that binds (10), when \( y \in (\bar{y}^{EE}, \bar{y}^{EL}) \). In this case, implicitly differentiating (10) yields

\[
\frac{d\rho}{dg} = \frac{1}{g} \left[ (1 - \delta \alpha f'_E(y)) \frac{dy}{dg} - \delta (1 - \alpha) f'_N(x) \frac{dx}{dg} - \rho \alpha \right].
\]

As \( dy/dg \leq 0 \) and \( dx/dg \geq 0 \), this implies \( d\rho/dg < 0 \). Finally, as \( \rho^* = 0 \) when \( y \leq \bar{y}^{EE} \) and \( \rho^* = 1 \) when \( y = \bar{y}^{EL} \), this implies that \( \rho^* \) is everywhere nonincreasing in \( g \).

**Proof of Proposition 7**

Note that, for all \( \gamma \geq \alpha \) and \( x \leq x^{FB} \), \( \gamma \)-weighted social welfare is increasing in \( x \), and increasing \( x \) relaxes (4). Hence, at the optimum either \( x^{EE}(\gamma) \geq x^{FB} \) or (3) binds. Let \( x^*(y) \) be the value of \( x \) that binds (3), and recall that the formula for \( dx^*(y)/dy \) is given by (14). Therefore, when \( x = x^*(y) \), the total derivative of social welfare with respect to \( y \) equals

\[
[(1 - \alpha) f'_N(x^*(y)) - (1 - \gamma)] \frac{dx^*(y)}{dy} + \alpha f'_E(y) - \gamma = \alpha f'_E(y) \left[ \frac{1 - \delta (1 - \gamma)}{1 - \delta (1 - \alpha) f'_N(x^*(y))} \right] - \gamma.
\]

Setting the derivative equal to 0 and rearranging yields

\[
\alpha f'_E(y) \left( \frac{\delta + \frac{1-\delta}{\gamma}}{\delta (1 - \alpha) f'_N(x^*(y))} \right) + \frac{\delta (1 - \alpha) f'_N(x^*(y))}{\delta (1 - \alpha) f'_E(y)} = 1.
\]

As the left-hand side of this equation is decreasing in \( y \), \( x^*(y) \), and \( \gamma \), and \( x^*(y) \) is nondecreasing in \( y \), it follows that the solution \( y \) (and hence \( x^*(y) \)) is nonincreasing in \( \gamma \).

Finally, since we have shown that \( x^{EE}(\gamma) = x^*(y) \) whenever \( x^{EE}(\gamma) < x^{FB} \), it follows that \( x^{EE}(\gamma) \) is nonincreasing in \( \gamma \) in a neighborhood of any \( \gamma \) such that \( x^{EE}(\gamma) < x^{FB} \). The fact that \( x^{EE}(\gamma) < x^{FB} \) then implies that \( x^{EE}(\tilde{\gamma}) \) (and hence \( y^{EE}(\tilde{\gamma}) \)) is nonincreasing on the entire interval \([\gamma', \gamma]\).

**Proof of Proposition 8**

The argument that \( x \) and \( y \) are nonincreasing in \( \gamma \) is the same as in Proposition 7. To show this implies that \( \rho^* \) is also nonincreasing, rewrite (10) as

\[
\rho^* g = (1 - \delta) y - \delta \left[ (1 - \alpha) f_N(x) + f_E(y) - \delta \bar{y}^g \right].
\]

Note that \( u_E \) is always nondecreasing in \( \gamma \). Hence, as \( y \) is nonincreasing, \( \rho^* \) is also nonincreasing.

**Proof of Proposition 9**

We first show that \( \frac{dy^*}{d\theta} \geq 0 \). Suppose instead that \( \frac{dy^*}{d\theta} < 0 \). We first show that this implies \( \frac{dx^*}{d\theta} \leq 0 \), and then show that \( \frac{dy^*}{d\theta} \) and \( \frac{dx^*}{d\theta} \) cannot both be negative.

As (3) binds at the optimum, \( x^*(\theta) = \delta \left[ (1 - \alpha) f_N(x^*(\theta), \theta) + \alpha f_E(y^*(\theta), \theta) \right] + g \).
To simplify notation, let \( f_N = f_N(x^*(\theta), \theta) \) and let \( f^E = f_E(y^*(\theta), \theta) \). Totally differentiating with respect to \( \theta \) yields
\[
\frac{dx^*}{d\theta} (1 - \delta (1 - \alpha) f^N_x) = \delta \left[ (1 - \alpha) f^N_\theta + \alpha f^E_y \frac{dy^*}{d\theta} + \alpha f^E_\theta \right]. \tag{15}
\]
Recall that \( 1 > \delta (1 - \alpha) f_N^x \) (because (3) binds and \( f_N \) is concave). Therefore, as \( f_N^N \) and \( f^E_\theta \) are non-positive, when \( \frac{dy^*}{d\theta} < 0 \), we also have \( \frac{dx^*}{d\theta} \leq 0 \).

Next, rewriting the first-order condition (11) using this notation, we have
\[
\alpha f_E^y + \delta (1 - \alpha) f_N^x = 1. \tag{16}
\]
Totally differentiating with respect to \( \theta \) yields
\[
\alpha f_E^y \frac{dy^*}{d\theta} + \alpha f_E^x \frac{dx^*}{d\theta} + \delta (1 - \alpha) f_N^x \frac{dx^*}{d\theta} + \delta (1 - \alpha) f_N^N = 0.
\]
As \( f_E^y \) and \( f_N^N \) are negative and \( f_N^x \) and \( f_E^x \) are non-negative, if \( \frac{dy^*}{d\theta} < 0 \) and \( \frac{dx^*}{d\theta} \leq 0 \) then we arrive at a contradiction. This establishes that \( \frac{dy^*}{d\theta} \geq 0 \).

It remains to show that \( \frac{dx^*}{dy^*} \geq 0 \). To see this, note that either \( \frac{dx^*}{dy^*} = 0 \) or \( \rho^* \in (0, 1) \). The former case is trivial. In the latter case, \( \rho^* \) is defined so as to bind the elites’ incentive constraint (10). That is,
\[
\rho^* (\theta) = \frac{1}{g} \left[ y^* (\theta) - \delta [(1 - \alpha) f_N (x^*(\theta), \theta) + \alpha f_E (y^*(\theta), \theta)] \right].
\]
Hence, \( \frac{dx^*}{dy^*} \) has the same sign as
\[
\frac{dy^*}{d\theta} - \delta \left( (1 - \alpha) f_N^N \frac{dx^*}{d\theta} + f^N_\theta \right) + \alpha \left( f_E^y \frac{dy^*}{d\theta} + f^E_\theta \right). \tag{17}
\]
Note that by (15),
\[
\frac{dx^*}{dy^*} = \frac{\delta \alpha f_E^y}{1 - \delta (1 - \alpha) f_N^x} + \frac{\delta [(1 - \alpha) f_N^N + \alpha f_N^E]}{1 - \delta (1 - \alpha) f_N^N} \leq \frac{\delta \alpha f_E^y}{1 - \delta (1 - \alpha) f_N^N}.
\]
Moreover, by (16),
\[
\frac{\delta \alpha f_E^y}{1 - \delta (1 - \alpha) f_N^N} = \delta.
\]
Hence, (17) equals
\[
\frac{dy^*}{d\theta} \left[ 1 - \delta \left( (1 - \alpha) f_N^N \frac{dx^*}{dy^*} + f^N_\theta \right) + \alpha \left( f_E^y + \frac{f_E^E}{dy^*/d\theta} \right) \right] \geq \frac{dy^*}{d\theta} \left[ 1 - \delta \left( (1 - \alpha) \delta f_N^N + \alpha f_E^y \right) \right]
= \frac{dy^*}{d\theta} \left[ 1 - \delta \right],
\]
where the last equation again follows by (16). Hence, \( \frac{dy^*}{d\theta} \geq 0 \) and \( \delta < 1 \) imply \( \frac{dx^*}{dy^*} \geq 0 \).
Proof of Proposition 10

The elites’ problem in this case becomes

\[
\max_{x \geq 0, y \in [0, \bar{y}]} (1 - \alpha) u_E (e_E + f_N (x)) + \alpha u_N (e_N + f_E (y)) - y
\]

subject to

\[
x \leq \delta [(1 - \alpha) u_N (e_N + f_N (x)) + \alpha u_N (e_N + f_E (y)) - u_N (e_N)] + g.
\]

Letting \( x^*(y) \) be the value of \( x \) that binds the constraint, we have

\[
\frac{dx^*}{dy} = \frac{\delta \alpha u_N' (e_N + f_N (x)) f'_N (y)}{1 - \delta (1 - \alpha) u_N' (e_N + f_N (x)) f'_N (x)}.
\]

With this equation, the elites’ first-order condition is

\[
(1 - \alpha) u_E'(e_E + f_N(x^*(y))) f'_N(x^*(y)) \frac{dx^*}{dy} + \alpha u_E'(e_E + f_E(y)) f'_E(y) = 1
\]

As \( x^*(y) \) is nondecreasing and \( u_E, f_N, \) and \( f_E \) are concave, we see that the left-hand side of the first-order condition is nonincreasing in both \( y \) and \( e_E \). Therefore, the optimal level of \( y \) (and hence the optimal level of \( x \)) is nonincreasing in \( e_E \). \( \blacksquare \)

Proof of Proposition 11

Imposing \( f_N = f_E = f \), we rewrite (12) as

\[
\max_{y \in [0, \bar{y}]} (1 - \alpha) f (x^*(y, \alpha)) + \alpha f (y) - y,
\]

where \( x^*(y, \alpha) \) is the value of \( x \) that makes (3) hold as equality when the fraction of elite agents is \( \alpha \). We now show that the solution to (18) is nonincreasing in \( \alpha \). Recall the relevant first-order condition in this case,

\[
\alpha f'(y) + \delta (1 - \alpha) f'(x^*(y, \alpha)) = 1.
\]

Implicitly differentiating yields

\[
\frac{dy}{d\alpha} = \frac{f'(y) - \delta f'(x^*(y, \alpha)) + \delta (1 - \alpha) f''(x^*(y, \alpha)) \frac{\partial x^*(y, \alpha)}{\partial x}}{\alpha f''(y) + \delta (1 - \alpha) f''(x) \frac{\partial x^*(y, \alpha)}{\partial y}}.
\]

Note that \( y \leq x^*(y, \alpha) \), and therefore \( f'(y) > \delta f'(x^*(y, \alpha)) \). In addition, \( x^*(y, \alpha) \) is nonincreasing in \( \alpha \) (again because \( y \leq x^*(y, \alpha) \)). Hence, the numerator in the above expression is positive and the denominator is negative, so the overall expression is positive. Hence, \( dy/d\alpha \geq 0 \), with strict inequality when \( y \) is interior.

Next, when \( y \in \left( \bar{y}^{EE}, \bar{y}^{EL} \right) \), and hence (10) binds, we have

\[
x^*(y, \alpha) = y + (1 - \rho) g.
\]

We may thus rewrite (10) as

\[
y = \delta [(1 - \alpha) f (y + (1 - \rho) g) + \alpha f (y)] + \rho g.
\]
Implicitly differentiating yields

\[
\frac{d\rho}{d\alpha} = \frac{1 - \delta \left( (1 - \alpha) f'(x^* (y, \alpha)) + \alpha f'(y) \right)}{g \left[ 1 - \delta (1 - \alpha) f'(x^* (y, \alpha)) \right]} + \delta \left[ f(x^* (y, \alpha)) - f(y) \right].
\]

In this expression, all three terms in brackets are positive. More specifically, the first is positive by the first-order condition; the second is non-negative as \( y \leq x^* (y, \alpha) \); and the third is positive by definition of \( x^* (y, \alpha) \). Hence, \( dy/d\alpha > 0 \) implies \( d\rho/d\alpha > 0 \). As \( \rho^* = 0 \) when \( y \leq \bar{y}_{EE} \) and \( \rho^* = 1 \) when \( y = \bar{y}_{EL} \), this implies that \( \rho^* \) is everywhere nonincreasing in \( g \).

**References**


Online Appendix: Proof of Proposition 12

In an equilibrium with effort levels \((w, x, y, z)\), expected per-period benefits of cooperation for a normal agent (gross of costs) are given by

\[
B_N(w, x, y, z) = (1 - \lambda\alpha) [(1 - \alpha)f_N(w) + \alpha f_E(y)] + \lambda\alpha [(1 - \alpha)f_N(x) + \alpha f_E(z)],
\]

and expected per-period benefits for an elite agent are given by

\[
B_E(w, x, y, z) = \lambda [(1 - \alpha)f_N(w) + \alpha f_E(y)] + (1 - \lambda (1 - \alpha)) [(1 - \alpha)f_N(x) + \alpha f_E(z)].
\]

The following lemma characterizes equilibria for a given level of equality before the law \(\rho\).

**Lemma 1** Given a level of equality before the law \(\rho\), there exists an equilibrium with effort levels \((w, x, y, z)\) if and only if

\[
(1 - \delta\alpha)w + \delta\alpha x \leq \delta B_N(w, x, y, z) + \delta\alpha g \tag{19}
\]

\[
(1 - \delta(1 - \alpha)) x + \delta(1 - \alpha) w \leq \delta B_N(w, x, y, z) + (1 - \delta(1 - \alpha)) g \tag{20}
\]

\[
(1 - \delta\alpha) y + \delta\alpha z \leq \delta B_E(w, x, y, z) + \delta\alpha pg \tag{21}
\]

\[
(1 - \delta(1 - \alpha)) z + \delta(1 - \alpha) y \leq \delta B_E(w, x, y, z) + (1 - \delta(1 - \alpha)) \rho g. \tag{22}
\]

**Proof.** In an equilibrium with effort levels \((w, x, y, z)\), we have

\[
(1 - \alpha)\mathbb{E}_{F_N}[f_N(x)] + \alpha\mathbb{E}_{F_E}[f_E(x)] = (1 - \alpha) [(1 - \alpha)f_N(w) + \alpha f_E(y)] + \alpha [(1 - \alpha)f_N(x) + \alpha f_E(z)].
\]

Hence, a normal agent’s equilibrium payoff is

\[
(1 - \lambda) [(1 - \alpha)f_N(w) + \alpha f_E(y)] + \lambda \mathbb{E}_{F_N}[f_N(x)] + \alpha\mathbb{E}_{F_E}[f_E(x)] - (1 - \alpha) w - \alpha x
\]

\[
= (1 - \lambda\alpha) [(1 - \alpha)f_N(w) + \alpha f_E(y)] + \lambda\alpha [(1 - \alpha)f_N(x) + \alpha f_E(z)] - (1 - \alpha) w - \alpha x,
\]

and elite agent’s equilibrium payoff is

\[
(1 - \lambda) [(1 - \alpha)f_N(x) + \alpha f_E(z)] + \lambda \mathbb{E}_{F_N}[f_N(x)] + \alpha\mathbb{E}_{F_E}[f_E(x)] - (1 - \alpha) y - \alpha z
\]

\[
= \lambda(1 - \alpha) [(1 - \alpha)f_N(w) + \alpha f_E(y)] + (1 - \lambda (1 - \alpha)) [(1 - \alpha)f_N(x) + \alpha f_E(z)] - (1 - \alpha) y - \alpha z.
\]

A normal agent’s incentive constraint when matched with another normal agent is thus

\[
(1 - \delta)(f_N(w) - w)
\]

\[
+ \delta ((1 - \lambda\alpha) [(1 - \alpha)f_N(w) + \alpha f_E(y)] + \lambda\alpha [(1 - \alpha)f_N(x) + \alpha f_E(z)] - (1 - \alpha) w - \alpha x)
\]

\[
\geq (1 - \delta)(f_N(w) - \delta\alpha g),
\]

where the left-hand side is a normal agent’s equilibrium payoff when matched with another normal agent and the right-hand side is a normal agent’s payoff from deviating to \(x_i = 0\) when matched with a normal agent and subsequently receiving her minmax payoff of \(-\alpha g\) (noting that a normal agent matched with another normal agent cannot be punished in the current period). This rearranges to (19). Similarly, a normal agent’s incentive constraint when matched with an elite is

\[
(1 - \delta)(f_E(y) - x)
\]

\[
+ \delta ((1 - \lambda\alpha) [(1 - \alpha)f_N(w) + \alpha f_E(y)] + \lambda\alpha [(1 - \alpha)f_N(x) + \alpha f_E(z)] - (1 - \alpha) w - \alpha x)
\]

\[
\geq (1 - \delta)(f_E(y) - g) - \delta\alpha g,
\]
as in this case a normal agent can be punished in the current period. This rearranges to (20). The argument for elite agents is similar, noting that an elite agent’s minmax payoff is \(-\rho a g\) rather than \(-\alpha g\).

Turning to the proof of the proposition, the elites’ problem is

\[
\max_{w,x,y,z,\lambda} \lambda (1-\lambda) \left[ (1-\lambda) f_N(w) + \alpha f_E(y) \right] + (1-\lambda) \left[ (1-\lambda) f_N(x) + \alpha f_E(z) \right] - (1-\lambda) y - \alpha z
\]

subject to (19)–(22). If \(\rho^* < 1\), then the elite incentive constraints (21) and (22) are slack, so the problem is equivalent to

\[
\max_{w,x,y,z} \lambda (1-\lambda) \left[ (1-\lambda) f_N(w) + \alpha f_E(y) \right] + (1-\lambda) \left[ (1-\lambda) f_N(x) + \alpha f_E(z) \right] - (1-\lambda) y - \alpha z
\]

subject to (19) and (20). We consider this less-constrained problem is what follows. In particular, we will show \(dw^*/dg \geq 0, dx^*/dg \geq 0, dy^*/dg \leq 0, dz^*/dg \leq 0\), and

\[
\begin{align*}
1 - \lambda \alpha f'_E(y^*) & \geq 0, \\
1 - (1-\lambda) f'_E(z^*) & \geq 0.
\end{align*}
\]

We first note that these inequalities imply \(d\rho^*/dg \leq 0\). To see this, recall that \(\rho^*\) is defined as the smallest value of \(\rho\) such that (21) or (22) binds. Implicitly differentiate (21) and (22) with respect to \(g\) to obtain

\[
\frac{dy^*}{dg} \left[ 1 - \delta \lambda - \delta \lambda (1-\lambda) \alpha f'_E(y^*) \right] + \frac{dz^*}{dg} \left[ \delta \lambda - \delta (1-\lambda) \alpha f'_E(z^*) \right] = \frac{dw^*}{dg} \left[ \delta \lambda (1-\lambda)^2 f'_N(w^*) \right] + \frac{dx^*}{dg} \left[ \delta (1-\lambda) (1-\lambda) f'_N(x^*) \right] + \delta \alpha \rho^* + \delta \alpha g \frac{d\rho^*}{dg}
\]

and

\[
\frac{dy^*}{dg} \left[ 1 - \delta \lambda - \delta \lambda (1-\lambda) \alpha f'_E(y^*) \right] + \frac{dz^*}{dg} \left[ 1 - \delta (1-\lambda) - \delta (1-\lambda) \alpha f'_E(z^*) \right] = \frac{dw^*}{dg} \left[ \delta \lambda (1-\lambda)^2 f'_N(w^*) \right] + \frac{dx^*}{dg} \left[ \delta (1-\lambda) (1-\lambda) f'_N(x^*) \right] + \delta \alpha \rho^* + \delta \alpha g \frac{d\rho^*}{dg}.
\]

Note that (23) and (24) imply that all bracketed terms in both of these equations are non-negative. Hence, if (23) and (24) hold, and in addition \(dw^*/dg \geq 0, dx^*/dg \geq 0, dy^*/dg \leq 0, dz^*/dg \leq 0\), then, whichever of (21) and (22) is the effective constraint, \(d\rho^*/dg\) must be non-positive.

To derive the desired inequalities, let \(u^E_{EL}(g)\) be the value of the elites’ problem for parameter \(g\). Note that \(u^E_{EL}(g)\) is a concave function of \(g\). To see this, suppose \((w,x)\) is a solution given coercive capacity \(g\) and \((w',x')\) is a solution given coercive capacity \(g' > g\). Then, for all \(\beta \in (0,1)\),

\[
(w^*,x^*) = (\beta w + (1-\beta) w', \beta x + (1-\beta) x')
\]

is feasible given coercive capacity \(\beta g + (1-\beta) g'\) (as \(f_N\) and \(f_E\) are concave), and elite utility at \((w^*,x^*)\) is greater than the \(\beta\)-weighted average of elite utility at \((w,x)\) and \((w',x')\).

Next, note that at least one of the normal agent incentive constraints (19) and (20) binds at the optimum. Suppose first that exactly one of these constraints binds. Letting \(\mu_{NN} \geq 0\) and \(\mu_{NE} \geq 0\) be the multipliers on (19) and (20), respectively,

\[
\frac{du^E_{EL}}{dg} = \delta \alpha \mu_{NN} + (1 - \delta (1-\lambda)) \mu_{NE}.
\]
As \( u^*_E (g) \) is concave, we also have
\[
\delta\alpha \frac{d\mu_{NN}}{dg} + (1 - \delta (1 - \alpha)) \frac{d\mu_{NE}}{dg} \leq 0.
\]

Since we have assumed that one of the two constraints binds, this implies that one of \( d\mu_{NN}/dg \) and \( d\mu_{NE}/dg \) is non-positive and the other is zero. Now, note that the first-order conditions in the less-constrained problem are given by
\[
\lambda (1 - \alpha)^2 f'_N (w) - \left[ \frac{\mu_{NN}}{\lambda} [1 - \delta \alpha - \delta (1 - \lambda \alpha) (1 - \alpha)] f'_N (w) \right] \left[ \frac{\mu_{NE}}{\lambda} [1 - \delta (1 - \alpha) - \delta \lambda \alpha (1 - \alpha)] f'_N (x) \right] = 0
\]
\[
(1 - \lambda (1 - \alpha)) (1 - \alpha) f'_N (x) - \left[ \frac{\mu_{NN}}{\lambda} [1 - \delta \alpha - \delta \lambda \alpha (1 - \alpha)] f'_N (x) \right] \left[ \frac{\mu_{NE}}{\lambda} [1 - \delta (1 - \alpha) - \delta \lambda \alpha (1 - \alpha)] f'_N (x) \right] = 0
\]
\[
\lambda (1 - \alpha) \alpha f'_E (y) - (1 - \alpha) \left[ \mu_{NN} + \mu_{NE} \right] \delta (1 - \lambda \alpha) \alpha f'_E (y) = 0,
\]
\[
(1 - \lambda (1 - \alpha)) \alpha f'_E (z) - \alpha \left[ \mu_{NN} + \mu_{NE} \right] \delta \lambda \alpha^2 f'_E (z) = 0.
\]

If (19) binds, then \( 1 - \delta \alpha - \delta (1 - \lambda \alpha) (1 - \alpha) f'_N (w) \geq 0 \) and \( \delta \alpha - \delta \lambda \alpha (1 - \alpha) f'_N (x) \geq 0 \), and if it is (20) that binds, then \( \delta (1 - \alpha) - \delta (1 - \lambda \alpha) (1 - \alpha) f'_N (w) \geq 0 \) and \( 1 - \delta (1 - \alpha) - \delta \lambda \alpha (1 - \alpha) f'_N (x) \geq 0 \) (otherwise, increasing \( w \) or \( x \) would relax the binding constraint while increasing the objective). As \( d\mu/dg \leq 0 \) for the binding constraint, the left-hand sides of the first two first-order conditions are nondecreasing in \( g \) for fixed \( w \) and \( z \). Hence, implicitly differentiating these first-order conditions with respect to \( g \) implies that \( dw^*/dg \) and \( dx^*/dg \) are both non-negative. Similarly, the left-hand sides of third and fourth first-order conditions are nonincreasing in \( g \) for fixed \( y \) and \( z \). Hence, implicitly differentiating these first-order conditions with respect to \( g \) implies that \( dy^*/dg \) and \( dz^*/dg \) are both non-positive. Finally, as the multipliers are non-negative, the third and fourth first-order conditions also yield
\[
1 - \alpha - \lambda (1 - \alpha) \alpha f'_E (y^*) \geq 0,
\]
\[
\alpha - (1 - \lambda (1 - \alpha)) \alpha f'_E (z^*) \geq 0.
\]

These inequalities imply (23) and (24), completing the proof in the case where exactly one of the normal agent constraints bind.

Finally, suppose that both (19) and (20) bind. In this case, \( g = x - w \), so substituting \( \delta \alpha (x - w) \) for \( \delta \alpha g \) in (19) and (20) lets us rewrite the elite’s problem as
\[
\max_{x,y,z} \lambda (1 - \alpha) [(1 - \alpha) f_N (x - g) + \alpha f_E (y)] + (1 - \lambda (1 - \alpha)) [(1 - \alpha) f_N (x) + \alpha f_E (z)] - (1 - \alpha) y - \alpha z
\]
subject to
\[
x = \delta [(1 - \lambda \alpha) [(1 - \alpha) f_N (x - g) + \alpha f_E (y)] + \lambda \alpha [(1 - \alpha) f_N (x) + \alpha f_E (z)]] + g.
\]
Let \( \mu_{NE} \geq 0 \) be the multiplier on (25). Then
\[
\frac{d\mu_{EL}}{dg} = \mu_{NE}.
\]
so the fact that $u^E_{EL}(g)$ is concave implies $d\mu_{NE}/dg \leq 0$. Finally the first-order conditions in the rewritten problem are

$$\begin{align*}
\lambda (1 - \alpha)^2 f'_N (x - g) + (1 - \lambda (1 - \alpha))(1 - \alpha) f'_N (x) + \mu_{NE} \left[ 1 - \delta (1 - \lambda \alpha)(1 - \alpha) f'_N (x - g) + \lambda \alpha (1 - \alpha) f'_N (x) \right] &= 0 \\
\lambda (1 - \alpha) \alpha f'_E (y) - (1 - \alpha) + \mu_{NE} \delta (1 - \lambda \alpha) \alpha f'_E (y) &= 0, \\
(1 - \lambda (1 - \alpha)) \alpha f'_E (z) - \alpha + \mu_{NE} \delta \lambda \alpha^2 f'_E (z) &= 0.
\end{align*}$$

By a similar argument as above, implicitly differentiating the first-order conditions with respect to $g$ yields $dx^*/dg \geq 0$ (and hence $dw^*/dg \geq 0$), $dy^*/dg \leq 0$, $dz^*/dg \leq 0$, (23), and (24). ■