A Theory of Equality Before the Law*

Daron Acemoglu
MIT

Alexander Wolitzky
MIT

September 11, 2020

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 elites 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 elites' perspective—to establish equality before the law, where all agents are subject to the same coercive punishments and norms are more equal. The key mechanism is that equality before the law increases elites' effort, which improves the carrot of future cooperation and thus encourages even higher effort from non-elites. Equality before the law combines high coercion and low inequality. Factors that make equality before the law more likely to emerge 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: rule of law, social norms, repeated games, community enforcement, coercion, inequality.
JEL Classification: D70, P16, P51, K10, C73.

*We thank Bob Gibbons, Gillian Hadfield, Rachel Kranton, Suresh Naidu, Jean Tirole, John Wallis, the anonymous referees, 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 argued 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]

We provide a simple framework for addressing this question. Our starting point is that society can be organized without a state, relying the threat of community enforcement or ostracism to support cooperative behavior, or it can be organized under the auspices of a state with the power to punish deviators, or equivalently to "enforce laws" coercively. State enforcement of laws not only affects incentives directly, but also influences 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 partial or full equality before the law (where laws apply more equally to all citizens), in each case with different implications for norms, incentives, and inequality. Our main objective is to shed light on the reasons societies 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, where "effort" stands for various pro-social behaviors, such as contributions to joint production,

[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 François, 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 elites voluntarily take steps towards equality before the law, though we also discuss how political pressures on the elite affect this choice, and compare these two mechanisms.

[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 breaking off cooperation still matter greatly even when legal enforcement is commonplace.

public goods, or collective defense. In a stateless society, effort (which we assume throughout is perfectly observed) can be enforced only by the carrot of continued societal cooperation. This can be achieved by standard community enforcement mechanisms, such as the threat of exclusion from cooperation, or ostracism. Though community enforcement can support positive equilibrium effort, this level is typically low, owing to the relatively 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 (who could be elites 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. Thus, in this "elite domination" equilibrium, all agents face the carrot of future cooperation, while normal agents are also confronted with the stick of coercive punishment.[3] Whereas under community enforcement there is relative equality across all agents, under elite domination the threat of punishment makes normal agents work much harder than elites, increasing inequality. Critically, this inequality is supported by both unequally applied laws and unequal norms—non-elite agents are coercively punished and suffer withdrawn cooperation when they do not exert the higher level of effort expected from them.[4] This unequal treatment implies that elites always benefit from coercion under elite domination, while normal agents may or may not benefit on net: they benefit from the higher effort of other normal agents, but suffer because they must 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 domination. 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 endogenize the extent to which elites are 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

[3] To be clear, the elite are above the law but are not "above norms": when they deviate from equilibrium behavior, they still suffer withdrawn 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. Moreover, social norms also regulate cooperation between elites.

[4] In principle, the legal effort standard ("law") that an agent must meet to avoid coercive punishment and the community effort standard ("norm") that she must meet to avoid community enforcement could be different from each other. However, we will see that in an efficient equilibrium these two standards always coincide.

make elites favor equality before the law.

Table 1 provides a schematic representation of the different enforcement regimes. (Our model is more complex that this schematic indicates, because pure community enforcement is formally a special case of elite domination where the coercive capacity of the state is zero, and there is a continuum of intermediate, "partial equality before the law" regimes in between elite domination and full equality before the law.) Pure community enforcement corresponds to low coercion and low inequality (but also low effort). Elite domination 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).[5]

	low coercion	high coercion
low inequality	pure community enforcement	equality before the law
high inequality	?	elite domination

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).[6] Finally, when the elites choose to transition to full equality before the law, they also voluntarily give up all of their privileges and exert the same level of effort as normal citizens, so there is full equality of norms as well as laws.[7]

What triggers the transition from elite domination to equality before the law? While our model highlights a number of factors affecting this tradeoff, we believe the most important one is the role of violence in society. We show that as the extent of coercive punishments that can be imposed on deviators decreases—for technological, political, or social reasons—it becomes more attractive for 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

[5] 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 (i.e., social demands and opportunities) may be associated with inequality of outcomes if skills or other attributes are unequally distributed. We also focus on equality between normal agents and elites, setting aside other possible legal inequalities, such as those among individuals in the same social or economic class (e.g., laws that advantage some workers over others).

[6] 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 domination.

[7] We emphasize that in our model elites do not value equality per se, but rather subject themselves to coercive punishments as a means of committing to higher future effort when normal agents also exert more effort. Because normal agents are initially subject to coercive punishments, this shift corresponds to greater equality before the law.

important and the carrot (the promise of future benefits) becomes more important.[8] This changes the tradeoff facing the elite and encourages them to increase their own effort. Finally, to achieve this increase in own effort, elites must subject themselves to coercive punishments.

This comparative static links our explanation for the emergence of equality before the law to political changes that strengthen non-elites and limit how harshly they can be treated by the state or the elite (cfr. footnote 1), as well as to social forces limiting 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 elites' political power declines, society again moves towards equality before the law.

Equality before the law can also emerge due to factors other than the diminished power of elites and limits on their ability to impose punishments. A notable possibility is that a change in the nature of production can alter the tradeoff 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 lenses of these two comparative static results, and we discuss several historical examples in Section 6. In particular, we illustrate how limits on coercive punishments can trigger the 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 the defensive modernization efforts in Japan during the Meiji Restoration (as well as several similar 19th-century reform movements), which are particularly interesting because they introduced elements of equality before the law even though they were driven by elites who did not face significant internal threats or bottom-up demands for fundamental political change, in contrast to the Athenian and British cases. Rather, such efforts illustrate how equality before the law helps mobilize society for national defense, modernization, and economic development.

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

[8] In other words, the stick and the carrot are substitutes. This is related to diminishing marginal returns to effort—the more effort is obtained by coercion, the less valuable is the marginal effort obtained by the threat of withdrawn cooperation.

devices for respecting other groups' property rights, and thus encouraging greater investment and economic participation.[9] 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, by removing elite privileges and increasing elite effort, encourages normal citizens to exert effort bears some resemblance 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).

The second literature is that on repeated games and community enforcement. Most of this literature focuses on the threat of withdrawing cooperation and does not consider costly punishments (Kandori, 1992, Ellison, 1994, Wolitzky, 2013, Ali and Miller, 2014).[10] A few papers do allow costly punishment, mostly focusing on enforcers' incentives to carry out punishments (Dixit, 2007, Masten and Prüfer, 2014, Levine and Modica 2016, Aldashev and Zanarone, 2017, Acemoglu and Wolitzky, 2019). These papers investigate neither enforcers' willingness to subject themselves to punishment nor equality before the law.

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 the model. Section 3 characterizes the best equilibrium for the elite under elite-domination, while Section 4 additionally characterizes the optimal degree of equality before the law from the viewpoint of the elite. Section 5

[9] 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).

[10] Kranton (1996) shows how access to formal markets can crowd out community enforcement. In our model, the presence of coercive legal enforcement strengthens community enforcement.

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 6. Section 7 discusses some possible extensions of our model, and Section 8 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 the withdrawal of cooperation and coercive punishment.

There is a continuum of infinitely-lived agents that discount the future with discount factor $\delta \in (0,1)$. Fraction α 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) $x_i \in \mathbb{R}_+$.[11] 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

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

Here, f_N and f_E are "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 latter assumption implies that $f'_N(0), f'_E(0) > 1$, so the social marginal benefit of effort initially exceeds its private marginal cost, which makes the stage game 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", or it could represent business investment 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. We also simplify the analysis by ruling out direct transfers between agents or groups.

Throughout we 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.[12]

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

[11] Effort 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.

[12] 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 note that our model is obviously the limit of a large finite population model. 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 α. 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.

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 ρg, where $\rho \in [0, 1]$ is a parameter measuring the vulnerability of elites to coercive punishment.[13]

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 society's values (is it socially acceptable to impose harsh punishments on law-breakers?). Meanwhile, the parameter ρ 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 withdrawing 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 withdrawing cooperation. Intermediate values of ρ in turn represent partial equality before the law. Such intermediate values may result in practice either because elites' 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).

In general, a strategy profile in this environment specifies, for each period, an effort level for each normal and elite agent as a function of all agents' past effort levels, as well as a set of agents to be punished at the end of the period as a function of all agents' past and current effort levels. 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.[14] As our main question is when elites themselves benefit from greater equality before the law, we focus on the best equilibrium for elites. That is, the equilibrium effort levels x and y are selected to maximize elites' payoffs, subject of course to incentive-compatibility constraints for both normal and elite agents. Not surprisingly, we will find that in an efficient equilibrium, agents who deviate from

[13] To be clear, we are modelling coercion as disutility that can be imposed on agents who deviate. Formally, each agent i remains "free" to take any action $x_i \in \mathbb{R}_+$.

[14] Non-stationary equilibria can potentially improve on stationary equilibria in discounted repeated games with 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".

their prescribed effort level will be punished by both the withdrawal of cooperation and coervice punishment (to the extend that they are vulnerable to such punishments). Thus, in an efficient equilibrium the prescribed effort levels x and y represent both norms (i.e., the standard of behavior required to avoid community punishment) and laws (i.e., the standard of behavior required to avoid legal punishment). A key aspect of our theory is that these standard can be different for normal and elite agents: that is, very often $x \neq y$.

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. In Appendix B, we analyze a more complex but closely related model where each period players randomly match in pairs, a player's effort disproportionately benefits her current partner, and players can only punish their partners. Our main results continue to apply in this alternative model.

3 Elite Domination

We first consider the case where $\rho = 0$, so elite agents are completely immune to coercion. We refer to this case as *elite domination*.

3.1 Elite-Optimal Equilibrium

Our first result characterizes the optimal equilibrium for elites when $\rho = 0$.

Proposition 1 *Under elite domination,*

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

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

$$x \leq \delta \left[(1-\alpha) f_N(x) + \alpha f_E(y) \right] + g, \tag{2}$$

$$y \leq \delta \left[(1-\alpha) f_N(x) + \alpha f_E(y) \right]. \tag{3}$$

2. *Constraint (2) binds at the optimum.*

3. *Let us denote the unique pair $(x, y) > (0, 0)$ such that both (2) and (3) bind by $\left(\bar{x}^{ED}, \bar{y}^{ED}\right)$, and denote the solution to (1) subject to (2) and (3) by $\left(x^{ED}, y^{ED}\right)$. Then we have*

$$\alpha f_E'\left(y^{ED}\right) + \delta(1-\alpha) f_N'\left(x^{ED}\right) \leq 1 \text{ if } y^{ED} = 0, \tag{4}$$

$$\alpha f_E'\left(y^{ED}\right) + \delta(1-\alpha) f_N'\left(x^{ED}\right) = 1 \text{ if } y^{ED} \in \left(0, \bar{y}^{ED}\right), \tag{5}$$

$$\alpha f_E'\left(y^{ED}\right) + \delta(1-\alpha) f_N'\left(x^{ED}\right) \geq 1 \text{ if } y^{ED} = \bar{y}^{ED}. \tag{6}$$

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

Second, (2) is the incentive constraint for a normal agent, while (3) is the incentive constraint for an elite agent.[15] To understand this, note that any agent (normal or elite) who deviates can lose the benefits of others' cooperation—worth $(1-\alpha) f_N(x) + \alpha f_E(y)$——in the next period. Moreover, normal agents who deviate face an additional coercive punishment of g. In contrast, when $\rho = 0$ there is no coercive punishment for elite agents, so this second term is not present in (3).

Specifically, the loss of the benefits of others' cooperation (which incentivizes effort by both normal and elite agents) can be supported by grim trigger strategies—in which cooperation completely breaks down following a deviation—combined with perpetual coercive punishment of normal agents who deviate. With these strategies, a normal agent's (per-period) equilibrium payoff is $(1-\alpha) f_N(x) + \alpha f_E(y) - x$, while her best payoff from deviating is $(1-\delta)[(1-\alpha) f_N(x) + \alpha f_E(y)] - g$. Equating the two yields (2). Similarly, an elite agent's equilibrium payoff is $(1-\alpha) f_N(x) + \alpha f_E(y) - y$, while her best payoff from deviating is $(1-\delta)[(1-\alpha) f_N(x) + \alpha f_E(y)]$. Equating these quantities yields (3). Thus, normal agents are punished by both community withdrawal of cooperation and coercive legal enforcement, while elites are punished only by withdrawal of cooperation. Note that in the best equilibrium for elites, normal agents are always required to work as hard as possible, so (2) binds.

While grim trigger strategies combined with the threat of perpetual punishment of normal agents are one way of supporting the unique optimal equilibrium effort level characterized in Proposition 1, they are not the only one. 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 χ_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 permanently ostracized. Introducing ostracism in this way does not affect our equilibrium conditions or results.[16]

Finally, (5) is the first-order condition with respect to y, once x has been substituted out of the objective function using (2). This expression captures the fact that elites benefit in two ways

[15] If we interpret y as the elite refraining from stealing and $f_E(y)$ as the harm that their extraction causes for normal agents, then (1) would need to be modified slightly by removing the $\alpha f_E(y)$ term from the objective function and the right-hand side of (3). This would have no major impact on our main results.

[16] 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).

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 community and legal enforcement, and is responsible for all of our comparative static results below. Since the indirect effect results from forward-looking behavior by normal agents, it vanishes 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 δ 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 $\delta^2(1-\alpha)^2 f'_N(x)^2$ units of benefit, and so on. The total marginal benefit to elites of increasing y is thus given by the geometric series

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

Equating this marginal benefit to the marginal cost of effort for the elite, which is 1, yields (5).[17]

In the next subsections, we discuss how the equilibrium described in Proposition 1 provides a stylized representation of social order in some historical elite-dominated societies. We start with the special case where $g = 0$—so coercive punishments are completely absent and all agents are incentivized only by the threat of withdrawn cooperation—and then consider the $g > 0$ case.

3.2 Stateless Societies under Community Enforcement: g = 0

When $g = 0$, Proposition 1, especially with the ostracism interpretation, provides a simple model of cooperation in stateless, small-scale societies. Note that some degree of pro-social behavior is supported despite the absence of 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. Instead, much violence in such

[17] 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

$$\alpha\left(f'_E(y)-1\right)+\delta(1-\alpha) f'_N(x)=1-\alpha,$$

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 Appendix B, we show that similar results obtain when effort creates a mix of public benefits and private returns for one's partner.

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 withdrawing 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 ostracism.[18] Moreover, as shown by Proposition 5 below, economic inequality is increasing in g, and hence attains its minimum value when $g = 0$. 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).

3.3 Limited Access Orders and Extractive Institutions: g>0

As g increases, the elite-optimal equilibrium features higher levels of inequality and coercion. Both of these are typical characteristics of early societies that developed state institutions (Johnson and Earle, 2000, Flannery and Marcus, 2014).[19] 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. Notably, in our model, unequal treatment before the law generates unequal norms as well: the threat of coercive legal punishment induces normal agents to exert higher effort, and normal agents who fail to deliver this greater 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 sharp distinction

[18] 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 $f'_N(x) > 1$ and $f'_E(y) > 1$ for effort levels that arise in equilibrium, it is more efficient to demand additional effort rather than transfers.

[19] An interesting question that our model does not address is how social hierarchy (i.e., an elite group, or $\alpha > 0$) and coercive capacity ($g > 0$) arise out of earlier egalitarian, small-scale societies. For recent economic models that speak to this question, see Dow and Reed (2013) and Mayshar, Moav, and Neeman (2017).

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 (1953, p. 98), "There were two great and universal divisions in the society...: between free men and serfs, and between free men and noblemen." Similarly, 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'." The laws that applied to lords were different than those that applied 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. Crucially, these norms, however, were unequal and privileged the elite in many domains (Bloch, 1939, Part VI).

3.4 The Effect of Coercion on Welfare

We now analyze how the welfare of normal and elite agents under elite domination depends on the coercive capacity of the state, g. It is clear that elites always benefit from a higher value of g, as this affects their maximization problem (1) only by relaxing the normal agent incentive constraint, (2). Normal agents face a tradeoff, however: an increase in g always increases reduces the share of the total surplus $(1-\alpha)(f_N(x)-x)+\alpha(f_E(y)-y)$ obtained by normal agents (i.e., increases inequality), but it can also increase the total surplus because the threat of coercive punishment increases the maximum sustainable effort level. The overall effect on normal agent welfare is ambiguous in general, but it can be characterized when the fraction of elite agents is small.

Proposition 2 *Assume $f_E'(0) < \infty$. There exists $\bar{\alpha} > 0$ such that if $\alpha < \bar{\alpha}$ then normal agent welfare in the elite-optimal equilibrium is single-peaked in g.*

The proof shows that when α is sufficiently small that elites do not find it in their interest to exert effort under elite domination, normal agent welfare is increasing in coercive capacity g up to a threshold $g^* \geq 0$ and is decreasing in g thereafter. The logic is that a small but positive level of g increases normal agent effort from the inefficiently low level arising at $g = 0$ towards the first-best level, but when g is too high normal agents are forced to exert effort above the first-best level. Thus, while the output-increasing effect of state enforcement can dominate if coercion is limited, the inequality effect dominates and normal agents are worse off when coercion is too intensive.

Proposition 2 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).[20] Either outcome is possible in our model: greater coercive capacity increases inequality, which makes normal agents worse off, but also increases total production, which benefits everyone.[21]

4 Endogenous Equality Before the Law

We now turn to our main focus: how equality before the law emerges endogenously. We thus now model elites' vulnerability to coercive punishment as a choice variable.

4.1 Elite-Optimal Equilibria

We now suppose that $\rho \in [0,1]$ is a choice variable for the elite, and continue to focus on the elite-optimal equilibrium.[22] The interpretation is that elites hold the political power to choose both the institutional environment (ρ) and the equilibrium. We assume that a cost $\varepsilon\rho$ is incurred by all agents when the extent of equality before the law is ρ. This may represent the cost of establishing the institutional foundations for greater equality before the law. In what follows, we take $\varepsilon \to 0$, which simplifies the analysis.[23] Under this assumption, the elites' problem becomes

$$\max_{x\geq 0, y\geq 0, \rho\in[0,1]} (1-\alpha) f_N(x) + \alpha f_E(y) - y \quad \text{s.t.} \tag{7}$$

$$x \leq \delta\left[(1-\alpha) f_N(x) + \alpha f_E(y)\right] + g, \tag{8}$$

$$y \leq \delta\left[(1-\alpha) f_N(x) + \alpha f_E(y)\right] + \rho g, \tag{9}$$

[20] To return to the example of medieval feudalism, the feudal order provided some security for commoners in an age when Europe was beset by warfare and suffered continual attacks from nomadic warriors; nevertheless, 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).

[21] Of course, elite agents always benefit from greater coercive capacity, which is also consistent with archaeological and historical evidence (e.g., Flannery and Marcus, 2014).

[22] In doing so, we also implicitly characterize the best equilibrium for the elite for any fixed value of ρ.

[23] The presence of this cost removes an uninteresting multiplicity. Since increasing ρ relaxes the incentive constraint of the elite and we focus on the elite-optimal equilibrium, if $\varepsilon = 0$ the elite would be willing to choose $\rho = 1$ (full equality before the law) and not punish themselves: intuitively, elites 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. Assuming $\varepsilon > 0$ rules out this artificial possibility and implies that the elites always choose the smallest level of ρ when indifferent. One could also obtain the same conclusion by introducing a small probability that elites suffer punishment on path and taking this probability to zero.

where (9) is the incentive compatibility constraint for elites, which must hold with equality if $\rho > 0$. Here (8) is identical to (2), while (9) differs from (3) in that an elite agent's minmax payoff is now $-\rho g$ rather than 0.

Let us denote the unique solution to the elites' problem—corresponding to the optimal equilibrium under endogenous equality before the law with minimal ρ—by $\left(x^{EL}, y^{EL}, \rho^*\right)$. 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 (8) to bind, as increasing x increases the objective and relaxes (9). Hence, $x^{EL} = x^*\left(y^{EL}\right)$, where again $x^*(y)$ is the value of x that satisfies (8) with equality. Let $\left(\bar{x}^{EL}, \bar{y}^{EL}\right)$ be the unique pair (x, y) such that $x = x^*(y)$ and (9) binds with $\rho = 1$. That is, $\left(\bar{x}^{EL}, \bar{y}^{EL}\right)$ are the greatest sustainable effort levels under full equality before the law. Note that $\bar{x}^{EL} = \bar{y}^{EL}$, so the maximum sustainable effort level for normal and elite agents is the same under full equality before the law.[24]

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

Proposition 3 *Every elite-optimal equilibrium takes one of the following three forms:*

1. *Elite domination:* $\rho^* = 0$, $\left(x^{EL}, y^{EL}\right) = \left(x^{ED}, y^{ED}\right)$, *and*

$$\alpha f_E'\left(\bar{y}^{ED}\right) + \delta(1-\alpha) f_N'\left(\bar{x}^{ED}\right) \leq 1.$$

2. *Partial equality before the law:* $\rho^* \in (0,1)$, $y^{EL} \in \left(\bar{y}^{ED}, \bar{y}^{EL}\right)$, $x^{EL} = x^*\left(y^{EL}\right) \in \left(\bar{x}^{ED}, \bar{x}^{EL}\right)$, *(9) binds, and*

$$\alpha f_E'\left(y^{EL}\right) + \delta(1-\alpha) f_N'\left(x^{EL}\right) = 1. \tag{10}$$

3. *Full equality before the law:* $\rho^* = 1$, $\left(x^{EL}, y^{EL}\right) = \left(\bar{x}^{EL}, \bar{y}^{EL}\right)$ *(in particular,* $x^{EL} = y^{EL}$*), and*

$$\alpha f_E'\left(\bar{y}^{EL}\right) + \delta(1-\alpha) f_N'\left(\bar{x}^{EL}\right) \geq 1.$$

The maximization problem (7) differs from (1) only in that ρ is now a choice variable, rather than being fixed exogenously at 0. As in the earlier problem, the incentive constraint for normal agents, (8), always binds, and that for elite agents, (9), binds only if the best equilibrium for elites involves maximum elite agent effort. Hence, if (9) with $\rho = 0$—or equivalently the corresponding constraint (3) under elite domination—is slack, then elites have no interest in committing themselves to a higher level of effort, and instead prefer to remain in the elite domination regime with $\rho = 0$. In contrast, if (3) binds under elite domination (or equivalently, if (6) holds with strict inequality),

[24] Note that 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 constraints, which are identical and thus imply the same level of effort.

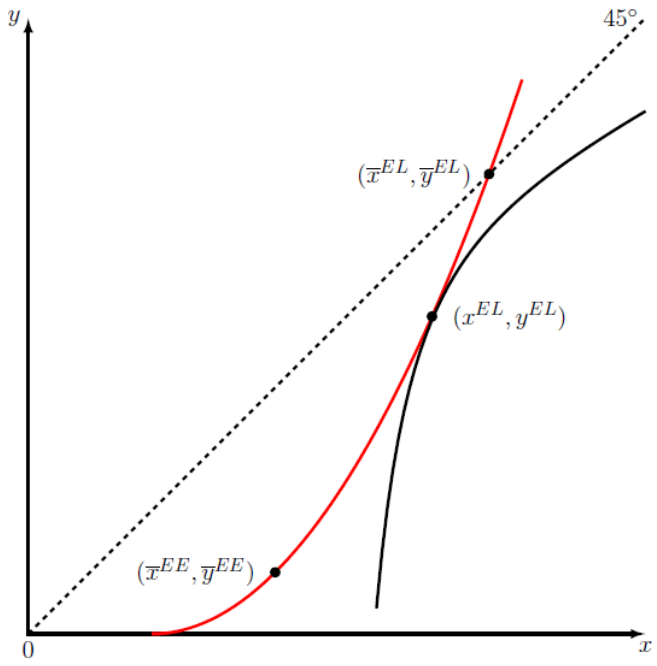

Figure 1: The black curve represents an indifference curve for the elite, while the red curve represents the boundary of the incentive compatibility constraint (8). The point $\left(\bar{x}^{EL}, \bar{y}^{EL}\right)$ corresponds to full equality before the law ($\rho = 1$) and the point $\left(\bar{x}^{EE}, \bar{y}^{EE}\right)$ corresponds to elite enforcement ($\rho = 0$).

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 (10).[25]

Finally, in the case where $\alpha f'_E\left(\bar{y}^{EL}\right) + \delta\left(1-\alpha\right) f'_N\left(\bar{x}^{EL}\right) \geq 1$, 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 economic equality (i.e., 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 3. Recall first that ρ^* is either 0 or the value of ρ that binds (9). We can thus omit (9) and rewrite the elites' problem, (7), as

$$\max_{x \geq 0, y \in [0, \bar{y}]} \left(1-\alpha\right) f_N\left(x\right) + \alpha f_E\left(y\right) - y \tag{11}$$

subject to (8), where $\bar{y} = \bar{y}^{ED}$ under elite domination and $\bar{y} = \bar{y}^{EL}$ under endogenous choice of ρ. We illustrate this problem in Figure 1. The red curve represents combinations of normal agent

[25] The intuition for this first-order condition with endogenous ρ is the same as for the one with $\rho = 0$ given in (5): the direct marginal benefit to elites of increasing their effort is $\alpha f'_E\left(y\right)$, and the indirect marginal benefit—coming through the induced increase in the maximum incentive compatible level of normal agent effort—is $\delta\left(1-\alpha\right) f'_N\left(x\right)$. The first-order condition sets the total marginal benefit of $\alpha f'_E\left(y\right) + \delta\left(1-\alpha\right) f'_N\left(x\right)$ equal to the total marginal cost of 1.

and elite effort that satisfy the normal agents' incentive constraint, (8), with equality. This curve intersects the 45° line at the point $\left(\bar{x}^{EL}, \bar{y}^{EL}\right)$, which corresponds to fully equality before the law, $\rho^* = 1$ (and equal effort from normal and elite agents). The point $\left(x^{ED}, y^{ED}\right)$, corresponding to elite domination with $\rho^* = 0$, is plotted as well. The figure also superimposes an indifference curve of (11). The tangency point, if any, between such an indifference curve and the boundary of (8) is the elite-optimal combination of (x, y); such a 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 $(x, y) = \left(\bar{x}^{EL}, \bar{y}^{EL}\right)$ with full equality before the law ($\rho^* = 1$), or at the point where $(x, y) = \left(x^{ED}, y^{ED}\right)$ with elite domination ($\rho^* = 0$).

4.2 Welfare

We next consider the implications of endogenous 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) level of equality before the law. Clearly, $u_E^{EL} \geq u_E^{ED}$, 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 4 *$u_N^{EL} \geq u_N^{ED}$, with strict equality if $\rho^* > 0$.*

That is, at the elite-optimal equilibrium with endogenous equality before the law, normal agents are always better-off than under elite domination. This follows because inequality is reduced and effort among all individuals is increased.

4.3 Interpretation

Like elite domination, partial or full equality before the law relies on the threat of coercive punishment to encourage pro-social behavior. However, it involves less inequality: elites' privileges are reduced or even completely eliminated under full equality before the law. 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 espoused by Hayek (1960), who emphasized the defining role of equal application of laws and equal protection from coercion. It is also 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 as in Acemoglu and Robinson (2012), which depend on a level economic playing field and thus the removal of various legal privileges. 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 additionally implies that, as exogenous parameters change such that the elite-optimal equilibrium transitions from elite domination to full equality before the law, it typically passes through a substantial region of partial equality before the law. This is consistent with the historical cases we discuss in Section 6, such as the Meiji Constitution in Japan and the *Tanzimat* reforms in the Ottoman Empire, which took important steps towards equality before the law but certainly did not result in full legal equality.

Finally, we have so far assumed that if the elite-optimal equilibrium involves some degree of equality before the law—$\rho > 0$—then elites can freely choose and commit to such an arrangement. An important question is how this can be secured in practice: that is, how can monopolization of coercion and enforcement 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, as the Meiji government did by disarming the samurai and creating a professional police force. A more modern version of this 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 judicial functions. We leave a more in-depth investigation of this important issue for future work.

5 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.

5.1 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 5 *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^{ED} and x^{EL} are strictly increasing in g, y^{ED} and y^{EL} are decreasing in g, and ρ^* is decreasing in g. In addition, if $\delta > 0$ and the solutions are interior, then the comparative statics on y^{ED}, y^{EL}, and ρ^* are strict.

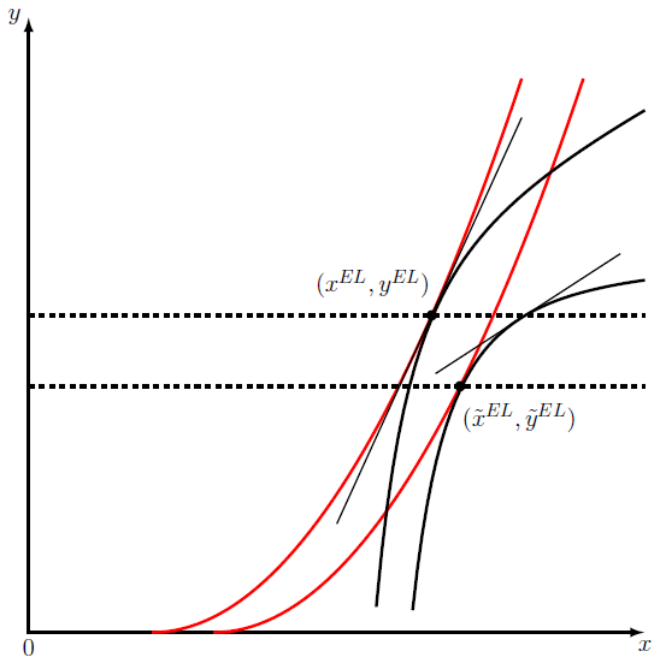

Figure 2: Elites' indifference curves become flatter as we move to the right along a horizontal line. An increase in g shifts the red curve representing the boundary of the normal-agent incentive constraint (8) to the right, and thus leads to a tangency point with greater x and lower y, and thus lower ρ^*.

Figure 2 illustrates the logic of this result. An increase in g has no impact on elites' indifference curves but shifts the boundary of (8) to the right. Elites' indifference curves become flatter as we move to the right along a horizontal line.[26] Consequently, shifting (8) leads to a new tangency 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 the threat of withdrawing co-operation are substitutes at the margin. The greater is g, the less "need" there is community enforcement, which lets elites reduce y. More precisely, recall that part of the elites' incentive to choose greater effort y is that this indirectly increases 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 (5), which captures this indirect effect, declines when x increases. This encourages the elite to choose a lower y. Since increasing ρ is a way to raise y (by making elites subject to greater coercive punishments), an increase in g also leads to lower ρ.[27] The fact that this comparative static ceases to be strict when $\delta = 0$ confirms this intuition, since in this case there

[26] This follows because the slope of the indifference curve is $-\frac{f'_N(x)}{1-f'_E(y)}$, which gets flatter as x increases (since f_N is concave).

[27] 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.

are no repeated game considerations and hence no indirect effect.[28]

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 19th-century Britain.

5.2 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) by simply considering the set of equilibria that maximize a weighted average of the utilities of elite and normal agents, and then reducing the weight on elites 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 elites.

Our first result establishes that under elite domination (or more generally, for any fixed level of equality before the law ρ), 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 intuition is that elites 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 6 *Under elite domination, let* $\left(x^{ED}(\gamma), y^{ED}(\gamma)\right)$ *denote the optimal equilibrium effort levels with Pareto weight* γ *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 (2) and (3). For all Pareto weights $\gamma > \gamma' \geq \alpha$*, if* $x^{ED}(\gamma) < x^{FB}$ *then* $x^{ED}(\gamma) \leq x^{ED}(\gamma')$ *and* $y^{ED}(\gamma) \leq y^{ED}(\gamma')$*.*

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 on elites 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 (8). 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.

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

Proposition 7 *Let* $\left(x^{EL}(\gamma), y^{EL}(\gamma), \rho^*(\gamma)\right)$ *denote the optimal equilibrium levels of effort and equality before the law with Pareto weight* γ *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 (8) and (9). 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* $\rho^*(\gamma) \leq \rho^*(\gamma')$*.*

5.3 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 (which induces greater equality before the law), while increasing average returns makes retaining their privileged position more attractive for elites. 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 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, ρ_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 θ. Assume f_N and f_E are twice continuously differentiable in (x, θ).

Proposition 8 *Suppose that increasing* θ *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,*

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

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

The comparative static 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^*.[29]

The result that $\frac{d\rho^*}{d\theta}$ is non-negative is subtle. Suppose increasing θ raises marginal returns while leaving average returns unchanged (a case allowed by the proposition). It is 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 already sufficient to justify the resulting higher level of elite effort? Intuitively, increasing θ 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 always falls 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.[30]

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 of our major results and will be discussed in detail in the next section.

5.4 Comparative Statics for the Size of the Elite

Our final comparative static says that a larger elite prefers a higher level of equality before the law. 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 9 *Assume $f_N = f_E = f$. Then an increase in the size of the elite, α, leads to an increase in elite agent effort and an increase in equality before the law. Formally, y^{ED}, y^{EL}, and*

[29] If we consider changes to the production functions that increase marginal returns at x_0 and y_0 without affecting average returns at these points—so that $\frac{\partial}{\partial\theta} f_N(x_0, \theta_0) = \frac{\partial}{\partial\theta} f_E(y_0, \theta_0) = 0$—then this offsetting effect is absent and hence $\frac{\partial x^*}{\partial\theta} \geq 0$. This may be a reasonable description of the changes that prompted the defensive modernizations we discuss.

[30] 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 (9) binds at ρ^*, we have

$$g\frac{d\rho^*}{d\theta} = \frac{dy^*}{d\theta} - \delta\frac{dB}{d\theta}.$$

At the elite-optimal equilibrium, we have $\frac{\partial B}{\partial y^*} = 1$. 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 out this argument in greater detail.

ρ^* are increasing in α. If the solutions are interior, then the comparative statics are strict.

To see the intuition, note that an increase in α 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 concave, the net effect is to raise the marginal benefit to elites of increasing y.[31] In contrast, the net effect of an increase in α on x is ambiguous, because the direct, negative effect may be offset by the indirect, positive effect coming through the increase in y.

6 Reinterpreting the Rise of Equality before the Law

In this section, we discuss several historical episodes 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.

6.1 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. Solon's reform agenda was acceptable to the elites partly because, as a wealthy merchant, he was one of them, and 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).

[31] The reason why this proposition, uniquely among our results, requires $f_N = f_E$ is that if $f'_E(y)$ is much smaller than $f'_N(x)$ even when $y \leq x$, then increasing α can decrease the net marginal benefit to elites of increasing y and reverse the comparative static. Note also that the comparative static with respect to α is strict even if $\delta = 0$, as changing α influences the direct effect term $\alpha f'\left(y^{EL}\right)$ in (5) in addition to the indirect effect.

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 owed 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. Though he came to power as the result of a mass uprising, Cleisthenes himself was from one of the most elite families in Athens. His reforms should therefore be viewed not as revolutionary but as attempts to strengthen Athens. Cleisthenes continued to build up Athenian judicial institutions and expanded state capacity by introducing various public services, financed for the first time by an elaborate fiscal system. Notably, many of his reforms were aimed at advancing equality before the law. Particularly important in this respect was the introduction of the ostracism law. 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, 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. First, though these reforms empowered regular Athenians, they were accepted and even embraced by elites. Second, they took place while the ability of elites to use coercion against regular Athenians was declining (Snodgrass, 1990, Ober, 2015). An important reason for this changing balance of power 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), "Cheap 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 make the elites themselves prefer and accept 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 a 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 different type of social conflict in Sparta.

6.2 Equality before the Law in Britain

The British case is often emphasized in discussions of the emergence 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).

One of the most important medieval English labor regulations was the Statute of Laborers, enacted in the 14th century, which empowered landowners to compel workers to work at set wages. In the words of the historian Robert Steinfeld, "The English laboring poor of this period... were subject to an oppressive regime of legal regulation" (2001, 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. Steinfeld writes, "in seventeenth-century England, the nearly universal legal form of consensual manual labor was not free labor but unfree labor" (p. 3). This regime was also exported to 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 dispro-

portionate power of employers over laborers less tenable. This process had started in the 17th century. Steinfeld notes, "It was during the seventeenth 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). These arguments began to be used to assert the rights of workers. For example, the Levellers argued in 1646 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 "the tone of English society changed" (p. 117). As a result, "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. Importantly, the 1867 Act was not meant to introduce equality before the law in relations between employers and workers. On the contrary, it simultaneously strengthened some aspects of the previous statutes and extended their coverage. However, consistent with our main comparative static highlighting the link between a reduction in the coercive power of elites and equality before the law, by taking coercive punishments off the table, this act marked the beginning of the end of unequal labor relations in Britain, as employers found their privileges harder to enforce. The Master and Servant 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. While the causes of the 1875 repeal are complex and undoubtedly include the changing values in British society that we have emphasized, the prior removal of employers' ability to use coercive punishments likely played an important role as well.

6.3 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. These cases also enable us to abstract from other sources of social change, such as citizens' demands for additional rights or elite responses to revolutionary threats. Though these factors have undoubtedly played a role in the historical evolution of institutions and can trigger the emergence of the equality before the law by limiting the extent of coercive punishments, they were not important in the context of defensive modernization episodes, such as the Meiji Restoration in 19th-century Japan. Such episodes thus elucidate how equality before the law may be useful for national defense, modernization, and economic development.

Japan in the first half of the 19th century was governed by the Tokugawa Shogunate, which established a highly hierarchical, class-based, quasi-feudal society. 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 roughly 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 (Macpherson, 1987; Jansen, 2002). These movements gained urgency and power with the threat posed by Perry's fleet, and ultimately convinced various factions of the Japanese elite to undertake an ambitious reform program as a defensive modernization strategy, with the goal of strengthening 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 Meiji. The Restoration was an elite-driven affair, a "coup organized by domain officials and court nobles" (Jansen, 2002, p. 336). Nevertheless, its objective was to initiate a process of military, social, and economic modernization, with the slogan *fukoku kyōhei* ("enrich the country, strengthen the army") (Macpherson, p. 24). 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.... [Now] Hereditary warriors were no longer needed. Instead, Japan's new military would comprise Japanese subjects from all classes" (2017, p. 5).

From the early stages, an emphasis on broader rights was part of the Meiji agenda. For example, the Charter Oath, issued by the young emperor in April 1868, included the following critical

provisions (Jansen, 2002, p. 338):

> "1. Deliberative assemblies shall be widely established and all matters decided by public discussion.
>
> 2. All classes, high and low, shall unite in vigorously carrying out the administration of affairs of state.
>
> 3. The common people, no less than the civil and military officials, shall each be allowed to pursue his own calling so that there may be no discontent.
>
> 4. Evil customs of the past shall be broken off and everything based upon the laws of Nature."

A number of important reforms followed. These included the 1871 lifting of the ban on intermarriage between commoners and samurai, the Registration Law empowering household heads with various freedoms, and the introduction of freedom of cultivation and legal land titles for farmers in the same year in 1873. In Ravina's summary, "The legal reforms of the caretaker government focused on rights, liberty, and individual autonomy," (2017, p. 147).

The watershed of these efforts was the Meiji Constitution, drafted in the 1880s and finally promulgated in 1890 (Jansen, 2002, pp. 365-66). 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. Even though there were relatively few bottom-up demands and no revolutionary threat from commoners at this time, the Meiji reforms reduced the standing of the elite (even if not that of the *daimyo* leaders). In Jansen's words, "Samurai surely experienced the greatest change of all, and most of them were clearly losers," (2002, p. 367).

An important question in this context is why the Meiji reforms did not simply modernize the military and the fiscal system, but also took major steps towards legal equality. As we noted already, this does not seem to have been driven by demands for such equality in Japanese society. Rather, consistent with the main mechanism and the second comparative static of our model, the trigger for these reforms—mounting external threats—increased the need for economic and military modernization. Moreover, the era's leaders decided that modernization could only be achieved by unifying the country, which necessitated the partial abolition of divisive elite privileges and stark inequalities. This is in line with Ravina's emphasis that these reforms were driven by the Restoration's leaders' beliefs that "connected military conscription with equality and liberty" (2017, p. 150) and viewed a limited form of equality before the law as essential for "creating a single, unified Japanese nation", capable of holding its own in international affairs (2017, p. 152).

6.4 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).

7 Extensions

We briefly discuss three possible extensions of our model.[32]

First, the model can be extended to analyze the effects of economic inequality between elite and normal agents. Suppose that, in addition to the public good $z = \alpha f_N(x) + (1-\alpha) f_E(y)$ produced by all agents' effort, each agent has a per-period endowment of consumption goods, which equals e_N for normal agents and e_E for elites. Suppose an agent's preferences over allocations of public and private goods are represented by an increasing, concave, and differentiable utility function $u(z, e)$ with a negative cross-partial derivative, so that public and private goods are substitutes. Then it can be shown that an increase in elites' endowments e_E leads to lower normal and elite agent effort, and hence reduces public good production. The intuition is that increasing e_E decreases elites' marginal utility from the public good, which reduces both the direct and indirect benefits of increasing y.[33] This observation that greater inequality in private endowments reduces public good

[32] More details are available in the working paper version of this paper (Acemoglu and Wolitzky, 2019).

[33] The impact on equality before the law is ambiguous, due to the offsetting effect that a higher endowment for

provision and leads in some cases to greater inequality before the law is consistent with several historical cases where early steps towards equality before the law were 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 Trefler, 2014).

It is also natural to extend our framework by introducing heterogeneity 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 economic and social modernization in late-medieval Western Europe resulted from the shifting political balance between different segments of the elite, in particular between commercial and landed interests (Moore, 1966, Aston and Philpin, 1987). We now describe how, in a simple extension 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.

Let us suppose that either (1) high-productivity elites' equilibrium effort is less than that of low-productivity elites, or (2) high-productivity elites benefit more than low-productivity elites from normal agent effort. There are many reasons why these conditions may hold. For example, if high-productivity elites can pretend to have low productivity (for example, an elite agent who controls both rural estates and factories can pretend that most of his income comes from estates, when it really mostly comes from factories), then members of the two elite subgroups cannot be asked to produce (very) different output levels. This then leads high-productivity elites to exert lower effort. Or, if elites' output levels are determined by intra-elite bargaining, we may expect to see relatively similar output levels between the two elite subgroups, which again leads high-productivity elites to exert lower effort. Finally, high-productivity elites may benefit disproportionately from normal agent effort if output is not a pure public good and normal agent effort (e.g. wage labor) is complementary with high-productivity elite effort (e.g. diligent operation of a factory).

When either condition (1) or (2) holds, a shift in political power from less productive to more productive elites (interpreted as an increase in the Pareto weight on more productive elites in a planning problem) increases all agents' effort and increases equality before the law. The reason is that, when high-productivity elites either exert less own effort or value normal agent effort more highly than do low-productivity elites, they are more inclined to trade off greater own effort for greater normal agent effort.

Elite agents may also be heterogeneous in terms of their position in the political or judicial hierarchy, which we can model as heterogeneity in their vulnerability to cercion. Suppose again that there are two elite groups, now corresponding to minor elites (say barons) and more powerful,

elites improves their payoffs under autarky, which makes a deviation more tempting. Greater equality before the law may then be required to counteract this heightened temptation to deviate. However, if the allocation of endowments is also socially determined—so that deviators can be stripped of their future-period endowments—then this second effect disappears. In this case, greater elite endowments unambiguously reduce equality before the law.

major elites (say dukes), which are 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 can be coerced by major elites, and the latter are initially completely immune to coercion. Finally, suppose that the level of equality before the law $\rho \in [0, 1]$ 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. Under reasonable conditions (spelled out in Acemoglu and Wolitzky, 2019), equality before the law increases when political power (i.e., Pareto weight) shifts towards the minor elites. 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 more than it increases that of minor elites, which makes minor elites more inclined to favor equality before the law.

This comparative static (like Proposition 9 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 individuals. 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 a group of English barons on King John in 1215, limiting his powers and ability to act without the baron's approval. 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 landowners' ability to impose forced labor on their own serfs (see Holt, 2015, and the discussion in Acemoglu and Robinson, 2019).

8 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, which relies 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 domination 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 having

immunity to coercion. Potentially shedding light on some major debates in anthropology, we show that the transition from community enforcement to elite domination (or more generally, increasing coercive capacity under elite domination) can increase or decrease the welfare of normal agents: productive effort increases, but so does inequality.

The most important part of our analysis concerns situations where the elite can choose between elite domination and various degrees of equality before the law. We show that it may be optimal—even from the elites' perspective—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 elites 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. Finally, consistent with arguments 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 the elite towards its weaker members—likewise favor greater equality before the law.

Many interesting areas remain to be explored. Possible further 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 also 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, through either repeated game incentives or gift-exchange type considerations.[34]

[34] A related direction would be to merge a model of labor coercion as in Acemoglu and Wolitzky (2011) with

Besides equality before the law, other aspects of the notion of rule of law include constraints on executive power—the sovereign must be bound by the law—and the primacy of law in conflict resolution. It may be fruitful to analyze the interaction of these features with equality before the law. Also, Hayek (1960) emphasizes the importance of the gradual evolution of the rule of law, an idea which is echoed by many legal philosophers, including Hart (1961); another important area for future research is to investigate the reasons for such gradual, evolutionary changes in the rule of law, and more generally the reasons for laws to be consistent with existing norms and customs. Finally, empirical analysis of the causes and implications of the emergence of equality before the law is another important area for subsequent research.

A Appendix: Proofs

A.1 Proof of Proposition 1

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

$$(1-\alpha) f_N(x) + \alpha f_E(y) - x \geq (1-\delta)\left[(1-\alpha) f_N(x) + \alpha f_E(y)\right] - g.$$

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 (2). The argument for (3) is the same, except that an elite agent's minmax payoff is 0 rather than $-g$. Moreover, (2) and (3) 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 (2) binds at the optimum, as increasing x increases the objective and also relaxes constraint (3).

For the last part of the result, let $x^*(y)$ be the value of x that binds (2). By the implicit function theorem,

$$\frac{dx^*(y)}{dy} = \frac{\delta\alpha f'_E(y)}{1-\delta(1-\alpha) f'_N(x^*(y))}.^{35} \tag{12}$$

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) $y > 0$, (3) is slack, and the derivative equals 0; or (iii) constraint (3) binds and the derivative is non-negative. This argument yields (4)–(6). ■

A.2 Proof of Proposition 2

As f_N is concave and $x^{ED} = \delta\left[(1-\alpha) f_N\left(x^{ED}\right) + \alpha f_E\left(y^{ED}\right)\right] + g$, we have that $\delta(1-\alpha) f'_N\left(x^{ED}\right) < 1$ uniformly over α. By (4)–(6) and $f'_E(0) < \infty$, there exists $\bar{\alpha} > 0$ such that if $\alpha < \bar{\alpha}$, then

repeated game considerations, so that the carrot of future cooperation interacts with coercive behavior by employers.

[35] The denominator is non-zero because, by concavity of f_N and inspection of (2), $1 - \delta(1-\alpha) f'_N(x)$ must be strictly positive at $x = x^*(y)$.

$y^{ED} = 0$ for all $g \geq 0$. Hence, for $\alpha < \bar{\alpha}$, $dx^{ED}/dg \geq 0$ (as x^{ED} is defined as the solution to $x = \delta(1-\alpha) f_N(x) + g$), and

$$\frac{du_N^{ED}}{dg} = \left((1-\alpha) f_N'\left(x^{ED}\right) - 1\right) \frac{dx^{ED}}{dg}.$$

So there exists $\hat{x}$ such that du_N^{ED}/dg is non-negative for $x^{ED} < \hat{x}$ and non-positive for $x^{ED} > \hat{x}$. Again using the fact that $dx^{ED}/dg \geq 0$, we conclude that u_N^{ED} is single-peaked in g. ■

A.3 Proof of Proposition 3

If the solution to the elites' problem involves $\rho^* = 0$, the problem reduces to that under elite domination. If instead $\rho^* > 0$, then (9) binds by the assumption that ρ^* is minimal. As (8) always binds, when $\rho^* = 1$ it immediately follows that $\left(x^{EL}, y^{EL}\right) = \left(\bar{x}^{EL}, \bar{y}^{EL}\right)$. When $\rho^* \in (0,1)$, the elite-optimal equilibrium is an interior solution to (7), subject to $y < \bar{y}^{EL}$. Hence, y^{EL} must satisfy the first-order condition (10) derived in the proof of Proposition 1. ■

A.4 Proof of Proposition 4

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

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

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

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

(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 $y^{EL} \geq y^{ED}$ is similar.

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

$$\begin{aligned} u_N^{ED} &= (1-\delta)\left[(1-\alpha) f_N\left(x^{ED}\right) + \alpha f_E\left(y^{ED}\right)\right] - g, \\ u_N^{EL} &= (1-\delta)\left[(1-\alpha) f_N\left(x^{EL}\right) + \alpha f_E\left(y^{EL}\right)\right] - g. \end{aligned}$$

As $x^{EL} \geq x^{ED}$, $y^{EL} \geq y^{ED}$, and f_N and f_E are increasing, it follows that $u_N^{EL} \geq u_N^{ED}$. ■

A.5 Proof of Proposition 5

Let $u_E(g)$ denote the value of (11) 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 β-weighted average of elite utility at (x, y) and (x', y').

Next, let μ_N be the Lagrange multiplier on (2). Note that

$$\frac{du_E(g)}{dg} = \mu_N.$$

Hence, μ_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

$$\begin{aligned} (1-\alpha) f_N'(x) &= \mu_N \left(1 - \delta(1-\alpha) f_N'(x)\right), \\ 1 - \alpha f_E'(y) &= \mu_N \delta \alpha f_E'(y). \end{aligned}$$

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 μ_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 ρ^*, recall that ρ^* is the value of ρ that binds (9), when $y \in \left(\bar{y}^{ED}, \bar{y}^{EL}\right)$. In this case, implicitly differentiating (9) yields

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

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

A.6 Proof of Proposition 6

Note that, for all $\gamma \geq \alpha$ and $x \leq x^{FB}$, γ-weighted social welfare is increasing in x, and increasing x relaxes (3). Hence, at the optimum either $x^{ED}(\gamma) \geq x^{FB}$ or (2) binds. Let $x^*(y)$ be the value of x that binds (2), and recall that the formula for $dx^*(y)/dy$ is given by (12). Therefore, when $x = x^*(y)$, the total derivative of social welfare with respect to y equals

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

Setting the derivative equal to 0 and rearranging yields

$$\alpha f_E'(y)\left(\delta + \frac{1-\delta}{\gamma}\right) + \delta(1-\alpha) f_N'(x^*(y)) = 1.$$

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

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

A.7 Proof of Proposition 7

The argument that x and y are nonincreasing in γ is the same as in Proposition 6. To show this implies that ρ^* is also nonincreasing, rewrite (9) as

$$\rho^* g = (1-\delta)\, y - \delta \underbrace{[(1-\alpha)\, f_N(x) + f_E(y) - \delta y]}_{=u_E}.$$

Note that u_E is always nondecreasing in γ. Hence, as y is nonincreasing, ρ^* is also nonincreasing. ■

A.8 Proof of Proposition 8

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 $\frac{dx^*}{d\theta}$ and $\frac{dy^*}{d\theta}$ cannot both be negative.

As (2) 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 θ yields

$$\frac{dx^*}{d\theta}\left(1 - \delta(1-\alpha) f_x^N\right) = \delta\left[(1-\alpha) f_\theta^N + \alpha f_y^E \frac{dy^*}{d\theta} + \alpha f_\theta^E\right]. \tag{13}$$

Recall that $1 > \delta(1-\alpha) f_x^N$ (because (2) binds and f_N is concave). Therefore, as f_θ^N and f_θ^E are non-positive, when $\frac{dy^*}{d\theta} < 0$, we also have $\frac{dx^*}{d\theta} \leq 0$.

Next, rewriting the first-order condition (10) using this notation, we have

$$\alpha f_y^E + \delta(1-\alpha) f_x^N = 1. \tag{14}$$

Totally differentiating with respect to θ yields

$$\alpha f_{yy}^E \frac{dy^*}{d\theta} + \alpha f_{y,\theta}^E + \delta(1-\alpha) f_{xx}^N \frac{dx^*}{d\theta} + \delta(1-\alpha) f_{x,\theta}^N = 0.$$

As f_{yy}^E and f_{xx}^N are negative and $f_{y,\theta}^E$ and $f_{x,\theta}$ 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 $\frac{d\rho^*}{d\theta} \geq 0$. To see this, note that either $\frac{d\rho^*}{d\theta} = 0$ or $\rho^* \in (0,1)$. The former case is trivial. In the latter case, ρ^* is defined so as to bind the elites' incentive constraint (9). That is,

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

Hence, $\frac{d\rho^*}{d\theta}$ has the same sign as

$$\frac{dy^*}{d\theta} - \delta\left[(1-\alpha)\left(f_x^N \frac{dx^*}{d\theta} + f_\theta^N\right) + \alpha\left(f_y^E \frac{dy^*}{d\theta} + f_\theta^E\right)\right]. \tag{15}$$

Note that by (13),

$$\begin{aligned}\frac{dx^*/d\theta}{dy^*/d\theta} &= \frac{\delta\alpha f_y^E}{1-\delta(1-\alpha)f_x^N} + \frac{\delta\left[(1-\alpha)f_\theta^N + \alpha f_\theta^E\right]}{1-\delta(1-\alpha)f_x^N} \\ &\leq \frac{\delta\alpha f_y^E}{1-\delta(1-\alpha)f_x^N}.\end{aligned}$$

Moreover, by (14),

$$\frac{\delta\alpha f_y^E}{1-\delta(1-\alpha)f_x^N} = \delta.$$

Hence, (15) equals

$$\begin{aligned}&\frac{dy^*}{d\theta}\left[1-\delta\left[(1-\alpha)\left(f_x^N\frac{dx^*/d\theta}{dy^*/d\theta} + \frac{f_\theta^N}{dy^*/d\theta}\right) + \alpha\left(f_y^E + \frac{f_\theta^E}{dy^*/d\theta}\right)\right]\right] \\ \geq\ &\frac{dy^*}{d\theta}\left[1-\delta\left[(1-\alpha)\delta f_x^N + \alpha f_y^E\right]\right] \\ =\ &\frac{dy^*}{d\theta}\left[1-\delta\right],\end{aligned}$$

where the last equation again follows by (14). Hence, $\frac{dy^*}{d\theta} \geq 0$ and $\delta < 1$ imply $\frac{d\rho^*}{d\theta} \geq 0$. ■

A.9 Proof of Proposition 9

Imposing $f_N = f_E = f$, we rewrite (11) as

$$\max_{y\in[0,\bar{y}]} (1-\alpha) f(x^*(y,\alpha)) + \alpha f(y) - y, \tag{16}$$

where $x^*(y,\alpha)$ is the value of x that makes (2) hold as equality when the fraction of elite agents is α. We now show that the solution to (16) is nonincreasing in α. Recall the relevant first-order condition in this case,

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

Implicitly differentiating yields

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

Note that $y \leq x^*(y,\alpha)$, and therefore $f'(y) > \delta f'(x^*(y,\alpha))$. In addition, $x^*(y,\alpha)$ is nonincreasing in α (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 $y \in \left(\bar{y}^{ED}, \bar{y}^{EL}\right)$, and hence (9) binds, we have

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

We may thus rewrite (9) as

$$y = \delta\left[(1-\alpha) f(y + (1-\rho) g) + \alpha f(y)\right] + \rho g.$$

Implicitly differentiating yields

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

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 $y \leq x^*\left(y,\alpha\right)$; and the third is positive by definition of $x^*\left(y,\alpha\right)$. Hence, $dy/d\alpha > 0$ implies $d\rho/d\alpha > 0$. As $\rho^* = 0$ when $y \leq \bar{y}^{ED}$ and $\rho^* = 1$ when $y = \bar{y}^{EL}$, this implies that ρ^* is everywhere nonincreasing in g. ■

B Decentralized Model with Private Benefits of Cooperation

Here we extend our model to one where agents match in pairs every period and effort disproportionately benefits one's current partner. Whereas in the pure public good model considered in the text elites can be favored only by having to exert less effort than normal agents, this richer model also allows them to benefit by receiving more effort from normal agents with whom they match.

Suppose that in every period, agents first randomly match in pairs and observe their partner's status (normal or elite) and then exert effort, and then each elite agent has the option of punishing her current partner. Assume 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/α on a normal agent and $\rho g/\alpha$ on an elite agent. (This scaling by $1/\alpha$ keeps the expected disutility of punishment fixed at g, as there are α elites in the population.) Assume also that a fraction $1-\lambda \in [0,1]$ of the benefits of cooperation accrue only to one's current 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 (as in the baseline model). 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

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

if her partner is normal, and is

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

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.

We show that our most important comparative static continues to hold in this model: increasing coercive capacity decreases equality before the law.[36]

Proposition 10 *Under endogenous equality before the law, suppose the elite-optimal level of equality before the law ρ^* 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 $d\rho^*/dg \leq 0$.*

[36] 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 certain changes in the environment. As a result, to be able to unambiguously sign these comparative statics, we would require additional assumptions, in particular conditions on third derivatives.

The basic intuition for this result is similar to that in the baseline model, in particular Proposition 5, 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 the baseline model, an increase in g relaxes normal agents' incentive 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 coercion.

References

[1] Abreu, Dilip. "Extremal Equilibria of Oligopolistic Supergames". *Journal of Economic Theory* 39 (1986): 191-225.

[2] Acemoglu, Daron, and Matthew O. Jackson. "Social Norms and the Enforcement of Laws". *Journal of the European Economic Association.* 15.2 (2017): 245-295.

[3] Acemoglu, Daron, and James A. Robinson. "Why Did the West Extend the Franchise? Democracy, Inequality, and Growth in Historical Perspective". *Quarterly Journal of Economics* 115.4 (2000): 1167-1199.

[4] Acemoglu, Daron, and James A. Robinson. *Economic Origins of Dictatorship and Democracy.* Cambridge University Press, 2005.

[5] Acemoglu, Daron, and James A. Robinson. *Why Nations Fail: The Origins of Power, Prosperity, and Poverty.* Crown, 2012.

[6] Acemoglu, Daron, and James A. Robinson. *The Narrow Corridor: States, Societies, and the Fate of Liberty*, Random House, New York, September 2019.

[7] Acemoglu, Daron, et al. "The Consequences of Radical Reform: The French Revolution". *American Economic Review* 101.7 (2011): 3286-3307.

[8] Acemoglu, Daron, and Alexander Wolitzky. "The Economics of Labor Coercion". *Econometrica* 72.9 (2011): 555-600.

[9] Acemoglu, Daron, and Alexander Wolitzky. "Sustaining Cooperation: Community Enforcement vs. Specialized Enforcement". *Journal of the European Economic Association* Forthcoming.

[10] Aldashev, Gani., Imane Chaara, Jean-Philippe Platteau, and Zki Wahhaj, "Using the Law to Change the Custom". *Journal of Development Economics* 97 (2012): 182-200.

[11] Aldashev, Gani, and Giorgio Zanarone. "Endogenous Enforcement Institutions". *Journal of Development Economics* 128 (2017): 49-64.

[12] Ali, S. Nageeb, and David A. Miller. "Enforcing Cooperation in Networked Societies". Working paper, 2014.

[13] Ali, S. Nageeb, and David A. Miller. "Ostracism and Forgiveness". *American Economic Review* 106.8 (2016): 2329-48.

[14] Aristotle. *The Constitution of Athens.* Translated by Frederic George Kenyon, Merchant Books, New York, 2009.

[15] Aston, Trevor Henry,and Charles HE Philpin, eds. *The Brenner Debate: Agrarian Class Structure and Economic Development in Pre-Industrial Europe.* Vol. 1. Cambridge University Press, 1987.

[16] Baker, George, Robert Gibbons, and Kevin J. Murphy. "Subjective performance measures in optimal incentive contracts". *Quarterly Journal of Economics* 109.4 (1994): 1125-1156.

[17] Baker, George, Robert Gibbons, and Kevin J. Murphy. "Relational Contracts and the Theory of the Firm". *Quarterly Journal of Economics* 117.1 (2002): 39-84.

[18] Bates, Robert H. *Prosperity and Violence: Political Economy of Development.* WW Norton, New York, 2001.

[19] Bates, Robert, Avner Greif, and Smita Singh. "Organizing violence". *Journal of Conflict Resolution* 46 (2002): 599-628.

[20] Baumard, Nicholas. "Has Punishment Played a Role in the Evolution of Cooperation? A Critical Review". *Mind and Society* 9 (2010): 171-192.

[21] Benabou, Roland, and Jean Tirole. "Laws and Norms". *Working Paper* (2011).

[22] Berman, Harold J. *Law and Revolution: The Formation of the Western Legal Tradition.* Harvard University Press, Cambridge, 1983.

[23] Besley, Timothy, and Torsten Persson. *Pillars of Prosperity: The Political Economics of Development Clusters.* Princeton University Press, 2011.

[24] Bidner, Chris, and Patrick Francois. "The Emergence of Political Accountability". *Quarterly Journal of Economics* 128.3 (2013): 1397-1448.

[25] Blanning, Timothy C. W. "The French Revolution and the Modernization of Germany". *Central European History* 22.2 (1989): 109-129.

[26] Bloch, Marc. *Feudal Society.* Two Volumes. University of Chicago Press, Chicago 9064.

[27] Boehm, Christopher. *Blood Revenge: The Enactment and Management of Conflict in Montenegro and Other Tribal Societies.* University of Pennsylvania Press, Philadelphia, 1986.

[28] Boehm, Christopher. *Hierarchy in the Forest: Egalitarianism and the Evolution of Human Altruism*, Harvard University Press, 1999.

[29] Bohannan, Paul and Laura Bohannan. *Tiv Economy.* Northwestern University Press, 1968.

[30] Briggs, Jean L. *Never in Anger: Portrait of an Eskimo Family.* Harvard University Press, 1970.

[31] Brenner, Robert. "Agrarian Class Structure and Economic Development in Pre-Industrial Europe". *Past & Present* 70 (1976): 30-75.

[32] Chagnon, Napoleon, *The Yanomamo.* Nelson Education, 1968.

[33] Childe, Gordon. *What Happened in History.* Penguin Books, London, 1942.

[34] Dixit, Avinash K. "Trade Expansion and Contract Enforcement". *Journal of Political Economy* 111 (2003): 1293-1317.

[35] Dixit, Avinash K. *Lawlessness and Economics: Alternative Modes of Governance*. Princeton University Press, 2007.

[36] Dow, Gregory K., and Clyde G. Reed. "The Origins of Inequality: Insiders, Outsiders, Elites, and Commoners," *Journal of Political Economy* 121.3 (2013): 609-641.

[37] Drew, Katherine Fischer. *The Laws of the Salian Franks*, University of Pennsylvania Press, 1991.

[38] Elias, Norbert. *The Civilizing Process*, Oxford: Blackwell 1994.

[39] Ellison, Glenn. "Cooperation in the Prisoner's Dilemma with Anonymous Random Matching". *Review of Economic Studies* 61 (1994): 567-588.

[40] Ember, Carol. "Myths about Hunter-Gatherers", *Ethnology*, 17, (1978): 439-448.

[41] Fearon, James D. "Self-Enforcing Democracy". *Quarterly Journal of Economics* 126.4 (2011): 1661-1708.

[42] Fisher, Herbert A. L. *Studies in Napoleonic Statesmanship: Germany*, Oxford; Clarendon Press, 1903.

[43] Flannery, Kent, and Joyce Marcus. *The Creation of Inequality: How our Prehistoric Ancestors Set the Stage for Monarchy, Slavery, and Empire.* Harvard University Press, 2014.

[44] Fukuyama, Francis. *The Origins of Political Order: From Prehuman Times to the French Revolution*. Farrar, Straus and Giroux, 2011.

[45] Gehlbach, Scott, and Philip Keefer. "Investment without Democracy: Ruling-Party Institutionalization and Credible Commitment in Autocracies". *Journal of Comparative Economics* 39.2 (2011): 123-139.

[46] Granovetter, Mark. "Economic Action and Social Structure: The Problem of Embeddedness". *American Journal of Sociology* 91.3 (1985): 481-510.

[47] Greif, Avner. "Contract Enforceability and Economic Institutions in Early Trade: The Maghribi Traders' Coalition". *American Economic Review* (1993): 525-548.

[48] Grossman, Herschel I. ""Make Us a King": Anarchy, Predation, and the State". *European Journal of Political Economy* 18.1 (2002): 31-46.

[49] Grossman, Sanford J., and Oliver D. Hart. "The Costs and Benefits of Ownership: A Theory of Vertical and Lateral Integration". *Journal of Political Economy* 94.4 (1986): 691-719.

[50] Hadfield, Gillian K., and Barry R. Weingast. "What is Law? A Coordination Model of the Characteristics of Legal Order" *Journal of Legal Analysis* 4.2 (2012): 471-514.

[51] Hart, H. L. A. *The Concept of Law*. Oxford (1961).

[52] Hart, Oliver, and John Moore. "Property Rights and the Nature of the Firm". *Journal of Political Economy* 98.6 (1990): 1119-1158.

[53] Hayek, Freidrich. *The Constitution of Liberty*, University of Chicago Press (1960).

[54] Holt, James C. *Magna Carta.* Third Edition, New York: Cambridge University Press, 2015.

[55] Huntington, Samuel. *Political Order in Changing Societies.* Yale University Press, New Haven, 1968.

[56] Jackson, Matthew O., and Yiqing Xing. "The Complementarity between Community and Government in Enforcing Norms and Contracts, and their Interaction with Religion and Corruption". *Working Paper* (2019).

[57] Jansen, Marius B. *The Making of Modern Japan.* Harvard University Press, Cambridge, MA 2002.

[58] Johnson, Allen W. and Timothy Earle. *The Evolution of Human Societies: From Foraging Group to Agrarian State.* Stanford University Press, 2000.

[59] Jones, Eric *The European Miralce: Environments, Economies and Geographies in the History of Europe and Asia.* New York: Cambridge University Press, 1981.

[60] Kandori, Michihiro. "Social Norms and Community Enforcement". *Review of Economic Studies* 59 (1992): 63-80.

[61] Knauft, Bruce. "Reconsidering Violence in Simple Human Societies", *Current Anthropology*, 28 (1987): 457-500.

[62] Konrad, Kai Andreas, and Stergios Skaperdas. 'The Market for Protection and the Origin of the State". *Economic Theory* 50 (2012): 417-443 .

[63] Kranton, Rachel E. "Reciprocal Exchange: A Self-Sustaining System". *American Economic Review* 84 (1996): 830-851.

[64] LeBlanc, Steven A., and Katherine E. Register. *Constant Battles: The Myth of the Peaceful, Noble Savae.* Macmillan, New York, 2003.

[65] Levi, Margaret. *Of Rule and Revenue.* University of California Press, 1989.

[66] Levine, David K., and Salvatore Modica. "Peer Discipline and Incentives Within Groups". *Journal of Economic Behavior & Organization* 123 (2016): 19-30.

[67] Lizzeri, Alessandro, and Nicola Persico. "Why Did the Elites Extend the Suffrage? Democracy and the Scope of Government, with an Application to Britain's "Age of Reform"". *Quarterly Journal of Economics* 119.2 (2004): 707-765.

[68] Macaulay, Stewart. "Non-Contractual Relations in Business: A Preliminary Study". *American Sociological Review* 28.1 (1963): 55-67.

[69] Macpherson, W. J. *The Economic Development of Japan, 1868-1941.* Cambridge University press, Cambridge 1995.

[70] Marlowe, Frank. *The Hadza: Hunter-Gatherers of Tanzania.* University of California Press, Berkeley 2010.

[71] Masten, Scott E., and Jens Prüfer. "On the Evolution of Collective Enforcement Institutions: Communities and Courts". *Journal of Legal Studies* 43 (2014): 359-400.

[72] Mayshar, Joram, Omer Moav, and Zvika Neeman. "Geography, Transparency, and Institutions". *American Political Science Review* 111.3 (2017): 622-636.

[73] Milgrom, Paul R., Douglass C. North, and Barry R. Weingast. "The Role of Institutions in the Revival of Trade: The Law Merchant, Private Judges, and the Champagne Fairs". *Economics & Politics* 2.1 (1990): 1-23.

[74] Moore, Barrington. *The Social Origins of Dictatorship and Democracy: Lord and Peasant in the Making of the Modern World.* Beacon Press, Boston, 1966.

[75] Moselle, Boaz, and Benjamin Polak. "A Model of a Predatory State". *Journal of Law, Economics, and Organization* 17.1 (2001): 1-33.

[76] Myerson, Roger B. "The Autocrat's Credibility Problem and Foundations of the Constitutional State". *American Political Science Review* 102.1 (2008): 125-139.

[77] Naidu, Suresh, and Noam Yuchtman. "Coercive Contract Enforcement: Law and the Labor Market in Nineteenth Century Industrial Britain". *American Economic Review* 103.1 (2013): 107-44.

[78] North, Douglass C., and Robert Paul Thomas. *The Rise of the Western World: A New Economic History.* Cambridge University Press, 1973.

[79] North, Douglass C., John Joseph Wallis and Barry R. Weingast. *Violence and Social Orders: A Conceptual Framework for Interpreting Recorded Human History.* Cambridge University Press, 2009.

[80] North, Douglass C., and Barry R. Weingast. "Constitutions and Commitment: the Evolution of Institutions Governing Public Choice in Seventeenth-Century England". *Journal of Economic History* 49.4 (1989): 803-832.

[81] Ober, Josiah. *The Rise and Fall of Classical Greece.* Penguin, New York, 2015.

[82] Osborne, Robin. *Greece in the Making, 1200-479 BC.* Rutledge, New York, 2009.

[83] Ostrom, Elinor. *Governing the Commons: The Evolution of Institutions for Collective Action.* Cambridge University Press, 1990.

[84] Owen, Roger. *State, Power and Politics in the Making of the Modern Middle East.* 3rd Edition, New York: Routledge, 2004.

[85] Pinker, Steven. *The Better Angels of Our Nature: Why Violence Has Declined.* Penguin, New York, 2011.

[86] Puga, Diego, and Daniel Trefler. "International Trade and Institutional Change: Medieval Venice's Response to Globalization". *Quarterly Journal of Economics* 129.2 (2014): 753-821.

[87] Radcliffe-Brown, Alfred R. *The Andaman Islanders.* Cambridge University Press, Cambridge, 2013.

[88] Ravina, Mark. *To Stand with the Nations of the World: Japan's Meiji Restoration in World History.* Oxford University Press, 2017.

[89] Rueschemeyer, Dietrich, Evelyne Huber Stephens, and John D. Stephens. *Capitalist Development and Democracy.* Polity: Cambridge, 1992.

[90] Sahlins, Marshall. *Stone Age Economics.* Routledge, London, 1972.

[91] Scott, James. *Against the Grain.* Yale University Press, New Haven, 2017.

[92] Snodgrass, Anthony M. *Archaic Greece: the Age of Experiment.* JM Dent, London, 1980.

[93] Southern, Richard. *The Making of the Middle Ages.* Yale University Press, 1953.

[94] Steinfeld, Robert J. *Coercion, Contract, and Free Labor in the Nineteenth Century.* Cambridge University Press, 2001.

[95] Suzman, James. *Affluence Without Abundance: The Disappearing World of the Bushmen.* Bloomsbury Publishing USA, 2017.

[96] Tyler, Tom R. *Why People Obey the Law.* Princeton University Press, 2006.

[97] Weingast, Barry R. "The Political Foundations of Democracy and the Rule of the Law". *American Political Science Review* 91.2 (1997): 245-263.

[98] Wiessner, Polly. "Norm Enforcement Among the Ju/'hoansi Bushmen". *Human Nature* 16.2 (2005): 115-145.

[99] Williamson, Oliver E. *Markets and Hierarchies: Analysis and Antitrust Implications.* Free Press, New York, 1975.

[100] Williamson, Oliver E. *The Economic Intstitutions of Capitalism.* Simon and Schuster, New York, 1985.

[101] Wolitzky, Alexander. "Cooperation with Network Monitoring". *Review of Economic Studies* 80 (2013): 395-427.

[102] Woodburn, James. "Egalitarian Societies". *Man* (1982): 431-451.

[103] Zürcher, Erik J. *Turkey: A Modern History.* IB Tauris, 2004.

C Online Appendix: Proof of Proposition 10

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

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

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

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

The following lemma characterizes equilibria for a given level of equality before the law ρ.

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

$$\begin{aligned}
(1-\delta\alpha) w + \delta\alpha x &\leq \delta B_N(w,x,y,z) + \delta\alpha g && (17)\\
(1-\delta(1-\alpha)) x + \delta(1-\alpha) w &\leq \delta B_N(w,x,y,z) + (1-\delta(1-\alpha)) g && (18)\\
(1-\delta\alpha) y + \delta\alpha z &\leq \delta B_E(w,x,y,z) + \delta\alpha\rho g && (19)\\
(1-\delta(1-\alpha)) z + \delta(1-\alpha) y &\leq \delta B_E(w,x,y,z) + (1-\delta(1-\alpha))\rho g. && (20)
\end{aligned}$$

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

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

Hence, a normal agent's equilibrium payoff is

$$\begin{aligned}
& (1-\lambda)\left[(1-\alpha) f_N(w) + \alpha f_E(y)\right] + \lambda\left[(1-\alpha)\mathbb{E}_{F_N}[f_N(x)] + \alpha\mathbb{E}_{F_E}[f_E(x)]\right] - (1-\alpha) w - \alpha x\\
= \;& (1-\lambda\alpha)\left[(1-\alpha) f_N(w) + \alpha f_E(y)\right] + \lambda\alpha\left[(1-\alpha) f_N(x) + \alpha f_E(z)\right] - (1-\alpha) w - \alpha x,
\end{aligned}$$

and elite agent's equilibrium payoff is

$$\begin{aligned}
& (1-\lambda)\left[(1-\alpha) f_N(x) + \alpha f_E(z)\right] + \lambda\left[(1-\alpha)\mathbb{E}_{F_N}[f_N(x)] + \alpha\mathbb{E}_{F_E}[f_E(x)]\right] - (1-\alpha) y - \alpha z\\
= \;& \lambda(1-\alpha)\left[(1-\alpha) f_N(w) + \alpha f_E(y)\right] + (1-\lambda(1-\alpha))\left[(1-\alpha) f_N(x) + \alpha f_E(z)\right] - (1-\alpha) y - \alpha z.
\end{aligned}$$

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

$$\begin{aligned}
& (1-\delta)\left(f_N(w) - w\right)\\
& +\delta\left[(1-\lambda\alpha)\left[(1-\alpha) f_N(w) + \alpha f_E(y)\right] + \lambda\alpha\left[(1-\alpha) f_N(x) + \alpha f_E(z)\right] - (1-\alpha) w - \alpha x\right]\\
\geq \;& (1-\delta) f_N(w) - \delta\alpha g,
\end{aligned}$$

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 (17). Similarly, a normal agent's incentive constraint when matched with an elite is

$$\begin{aligned}
& (1-\delta)\left(f_E(y) - x\right)\\
& +\delta\left[(1-\lambda\alpha)\left[(1-\alpha) f_N(w) + \alpha f_E(y)\right] + \lambda\alpha\left[(1-\alpha) f_N(x) + \alpha f_E(z)\right] - (1-\alpha) w - \alpha x\right]\\
\geq \;& (1-\delta)\left(f_E(y) - g\right) - \delta\alpha g,
\end{aligned}$$

as in this case a normal agent can be punished in the current period. This rearranges to (18). The argument for elite agents is similar, noting that an elite agent's minmax payoff is $-\rho\alpha g$ rather than $-\alpha g$. ■

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

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

subject to (17)–(20). If $\rho^* < 1$, then the elite incentive constraints (19) and (20) are slack, so the problem is equivalent to

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

subject to (17) and (18). 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

$$1 - \lambda\alpha f_E'(y^*) \geq 0, \tag{21}$$
$$1 - (1-\lambda(1-\alpha)) f_E'(z^*) \geq 0. \tag{22}$$

We first note that these inequalities imply $d\rho^*/dg \leq 0$. To see this, recall that ρ^* is defined as the smallest value of ρ such that (19) or (20) binds. Implicitly differentiate (19) and (20) with respect to g to obtain

$$\begin{aligned} &\frac{dy^*}{dg}\left[1 - \delta\alpha - \delta\lambda(1-\alpha)\alpha f_E'(y^*)\right] + \frac{dz^*}{dg}\left[\delta\alpha - \delta(1-\lambda(1-\alpha))\alpha f_E'(z^*)\right] \\ = \;& \frac{dw^*}{dg}\left[\delta\lambda(1-\alpha)^2 f_N'(w^*)\right] + \frac{dx^*}{dg}\left[\delta(1-\lambda(1-\alpha))(1-\alpha) f_N'(x^*)\right] + \delta\alpha\rho^* + \delta\alpha g\frac{d\rho^*}{dg} \end{aligned}$$

and

$$\begin{aligned} &\frac{dy^*}{dg}\left[\delta(1-\alpha) - \delta\lambda(1-\alpha)\alpha f_E'(y^*)\right] + \frac{dz^*}{dg}\left[1 - \delta(1-\alpha) - \delta(1-\lambda(1-\alpha))\alpha f_E'(z^*)\right] \\ = \;& \frac{dw^*}{dg}\left[\delta\lambda(1-\alpha)^2 f_N'(w^*)\right] + \frac{dx^*}{dg}\left[\delta(1-\lambda(1-\alpha))(1-\alpha) f_N'(x^*)\right] + \delta\alpha\rho^* + \delta\alpha g\frac{d\rho^*}{dg}. \end{aligned}$$

Note that (21) and (22) imply that all bracketed terms in both of these equations are non-negative. Hence, if (21) and (22) hold, and in addition $dw^*/dg \geq 0$, $dx^*/dg \geq 0$, $dy^*/dg \leq 0$, and $dz^*/dg \leq 0$, then, whichever of (19) and (20) 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 β-weighted average of elite utility at (w, x) and (w', x').

Next, note that at least one of the normal agent incentive constraints (17) and (18) 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 (17) and (18), respectively,

$$\frac{du_E^{EL}}{dg} = \delta\alpha\mu_{NN} + (1-\delta(1-\alpha))\mu_{NE}.$$

As $u_E^{EL}(g)$ 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

$$\begin{aligned}
\lambda(1-\alpha)^2 f_N'(w)-\left[\begin{array}{c}\mu_{NN}\left[1-\delta\alpha-\delta(1-\lambda\alpha)(1-\alpha)f_N'(w)\right]\\+\mu_{NE}\left[\delta(1-\alpha)-\delta(1-\lambda\alpha)(1-\alpha)f_N'(w)\right]\end{array}\right] &= 0\\
(1-\lambda(1-\alpha))(1-\alpha)f_N'(x)-\left[\begin{array}{c}\mu_{NN}\left[\delta\alpha-\delta\lambda\alpha(1-\alpha)f_N'(x)\right]\\+\mu_{NE}\left[1-\delta(1-\alpha)-\delta\lambda\alpha(1-\alpha)f_N'(x)\right]\end{array}\right] &= 0\\
\lambda(1-\alpha)\alpha f_E'(y)-(1-\alpha)+(\mu_{NN}+\mu_{NE})\delta(1-\lambda\alpha)\alpha f_E'(y) &= 0,\\
(1-\lambda(1-\alpha))\alpha f_E'(z)-\alpha+(\mu_{NN}+\mu_{NE})\delta\lambda\alpha^2 f_E'(z) &= 0.
\end{aligned}$$

If (17) 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 (18) 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

$$\begin{aligned}
1-\alpha-\lambda(1-\alpha)\alpha f_E'(y^*) &\geq 0,\\
\alpha-(1-\lambda(1-\alpha))\alpha f_E'(z^*) &\geq 0.
\end{aligned}$$

These inequalities imply (21) and (22), completing the proof in the case where exactly one of the normal agent incentive constraints bind.

Finally, suppose that both (17) and (18) bind. In this case, $g=x-w$, so substituting $\delta\alpha(x-w)$ for $\delta\alpha g$ in (17) and (18) lets us rewrite the elite's problem as

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

subject to

$$x=\delta\left[(1-\lambda\alpha)\left[(1-\alpha)f_N(x-g)+\alpha f_E(y)\right]+\lambda\alpha\left[(1-\alpha)f_N(x)+\alpha f_E(z)\right]\right]+g. \qquad (23)$$

Let $\mu_{NE}\geq 0$ be the multiplier on (23). Then

$$\frac{du_E^{EL}}{dg}=\mu_{NE},$$

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

$$\begin{aligned}
\left[\begin{array}{c} \lambda(1-\alpha)^2 f_N'(x-g) \\ +(1-\lambda(1-\alpha))(1-\alpha) f_N'(x) \end{array}\right] + \mu_{NE}\left[\begin{array}{c} 1-\delta(1-\lambda\alpha)(1-\alpha) f_N'(x-g) \\ +\lambda\alpha(1-\alpha) f_N'(x) \end{array}\right] &= 0 \\
\lambda(1-\alpha)\alpha f_E'(y) - (1-\alpha) + \mu_{NE}\delta(1-\lambda\alpha)\alpha f_E'(y) &= 0, \\
(1-\lambda(1-\alpha))\alpha f_E'(z) - \alpha + \mu_{NE}\delta\lambda\alpha^2 f_E'(z) &= 0.
\end{aligned}$$

By a similar argument as above, implicitly differentiating the first-order conditions with respect to g yields $dx^*/dg \geq 0$ (and hence $dw^*/dg \geq 0$), $dy^*/dg \leq 0$, $dz^*/dg \leq 0$, (21), and (22). ■